INTERNATIONAL HELLENIC UNIVERSITY

Constructing Homeland Security Knowledge Graphs

Emmanouil Koulas

SID: 3308190013

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

APRIL 2021

THESSALONIKI – GREECE

Constructing Homeland Security Knowledge Graphs

Emmanouil Koulas

SID: 3308190013

Supervisor:	Christos Berberidis
Supervising Committee Members:	Leonida Akritidis
	Apostolos Papadopoulos

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

APRIL 2021

THESSALONIKI – GREECE

Abstract

This dissertation was written as a part of the MSc in Data Science at the International Hellenic University. This project is placed on the intersection of traditional security domains, like homeland security, and technology. The importance of incorporating new technologies to address traditional threat to security is almost self-evident.

In this dissertation I construct two knowledge graphs related to terrorism, using the RAND Database of WorldWide Terrorism Incidents (RANDWTI) and the Global Terrorism Database (GTD). Utilizing Neo4J's software to construct the graphs, I highlight some use cases in which the very existence of those graphs can provide valuable insight to policy makers and law enforcement agencies alike. The rise of computational power combined with the availability of big, case-specific, data will enable wide deployment of such initiatives.

I would like to thank my supervisor, Dr. Christos Berberidis, for the inspiration to join and complete the master's programme, Mrs Eleni Kapantai, for her invaluable support during the early stages of the thesis, Mrs Chrysothea Basia, for her guidance and Mrs Stella Dimitsaki for her technical support. Finally, I would like to thank my parents, Antonios and Nikoletta, my brother Efthymios for all the late-night discussions, and my grandmother Efthymia who was more eager for our academic advancement than we ever were, who passed away before having the chance of seeing the dissertation completed.

Emmanouil Koulas

29 April 2021

Contents

List of Figures

List of Tables

1 Introduction

The IT sector has evolved rapidly in recent decades resulting in the digitization of processes and databases of both private and public organizations. The effects of the proliferation of internet use have both positive and negative effects on the general population as well as organizations and the states across the globe (Brass & Sowell, 2020; Gruenewald, Allison-Gruenewald, & Klein, 2015; Kwon & Rao, 2017; Kelarestaghi, Heaslip, Khalilikhah, Fuentes, & Fessman, 2018; Li, Xu, Wang, Chen, & Sun, 2018; Papadaki et al., 2017; Noel, Harley, Tam, Limiero, & Share, 2016).

The positives include the automation of a variety of processes, the immediacy of communication and information as well as the ease of access to databases and files. Also, due to the spread of the internet and the evolution of information technologies, new opportunities have been created for the economy, production and industry as well as the organization of various functions. The negatives include the increase in threats, especially in terms of network security, information and users' personal data. However, the threats seem to be increasing as well as the possibilities for dealing with them (Awan & Memon, 2016; Gardner, 2014; Koulas, 2019, Koulas et al., 2020).

At the state level, different strategies are developed and different programs are implemented which aim to facilitate the authorities in maintaining the security of citizens and their national interests. Typically, the European Union (hereinafter EU), the United States of America (hereinafter the USA) but also Asian countries such as China, Japan and India, participate in programs to address asymmetric threats (terrorism, international crime) and of simple breaches in their networks (Awan & Memon, 2016; Gardner, 2014; Gruenewald, Allison-Gruenewald, & Klein, 2015; Jung & Park, 2014). For the EU, in particular, it is observed that this task is undertaken by the Union itself in coop-

eration with NATO and, at national level, by its Member States separately (Koulas, 2019).[1]

Knowledge graphs are a state-of-the-art assistive technology that was originally developed by software companies to help users search for information and navigate web pages (Pirrò, 2015). For example, Google, Yahoo, and Facebook, like many other companies, use knowledge graphs to correlate information and facilitate search. The logic of knowledge graphs is to make connections between information, people and events and to reveal relationships, for example, between two people (Noy, et al., 2019).

Typically, knowledge graph technology can be used to analyze whether a terrorist attack is linked to a person, or whether two terrorists are linked. These technologies are so highly developed that this information can be linked and patterns created even if their relationship is not immediately apparent. Also, knowledge graphs work because they use both material that has already been categorized in databases (tables) and material that is related to the lower case, for example various texts that are published on the internet (Pirrò, 2015).

As a technology, knowledge graphs perform in various fields and offer the opportunity for immediate processing and disclosure of relationships between seemingly unrelated factors. In cybersecurity, for example, knowledge graphs can provide information to the user about the threat of cyber intrusions, the security of users and information, and the type of threats (Collarana, et al., 2018; Iannacone, et al., 2015; Noy, et al., 2019).

---

[1] In the United States in particular, however, there seems to be an active dialogue between political leaders, pressure groups, the media and public opinion, especially after 9/11, regarding the state's ability to secure borders and the cyberspace of the United States. The main reason is the attack on the Twin Towers in 2001, which is understood by internationalists as the day that finally marked the end of the short period of US hegemony in the International System. Moreover, in the United States, there is a strong emphasis on tackling the refugee issue and illegal immigration, as well as early detection of international threats (Mearsheimer, 2011).

Also, the application of knowledge graphs so far in homeland security has proven to be important for the following reasons (Kaynar, 2016; Miehling, Rasouli, & Teneketzis, 2015; Noel, Harley, Tam, Limiero, & Share, 2016):

- They can give the users perspective and information

- To bring to the surface patterns that will help the states to deal more effectively with the threats that may concern the public and national security and the protection of the national interests of the states.

This paper will examine the issue of Homeland Security with an emphasis on the strategies of the USA, the EU and Asian countries such as China and India for tackling terrorism, international crime and illegal immigration. Moreover, the scope of this paper is to identify the challenges associated with terrorism in the aforementioned regions using knowledge graphs.

In detail, the paper will be structured as such:

- The first chapter is the introductory one
- The second chapter consists of the literature review. The methodology used is that of the systematic literature review assessing a wide variety of articles that refer to either national security and the strategies of the regions under study or the use of knowledge graphs and attack graphs in the study of terrorism and its impact in those regions.
- The third chapter is that of the methodology of the research part of the paper, meaning the transformation of data with regard to terrorism, and the ontology for the purpose of creating a knowledge graph.
- The fourth chapter focuses on the presentation of the data and findings. In this chapter some testing queries are run, in order to show the capabilities of the knowledge graph.
- Last, the final chapter, that of the conclusions, is dedicated to a summary of the findings of the paper as well as the discussion of the limitations of the present study as well as the author's suggestions for future research.

The contribution of this research to the available academic literature is great. This is because homeland security is extremely important in ensuring the well-being of states

and the use of knowledge graphs can go a long way in maintaining an optimal level of security for states around the world.

However, due to the fact that this thesis revolves around a rather contested topic, it is crucial to underline that the opinions, statements and findings of the paper do not, necessarily, reflect the views of the supervisors of the thesis. Accordingly, any statements and comments are not, necessarily, supported by either the supervisors or any other members of the University and are conclusions drawn by the study of material as well as the modelling of the knowledge graphs.

2 Literature Review

In this chapter I will review relevant literature, as well as, the methodology used to acquire said literature.

2.1 Methodology

A systematic and / or critical literature review are amongst the most used methods of elaboration of studies and works and contributes to ensuring a broader, global and absolute understanding of the subject. In the present dissertation, the method of systematic literature review (S.L.R.) was chosen for the theoretical study of the two main topics related the work. These are:

1. homeland security (HS), and
2. the use of knowledge graphs for HS purposes.

For the theoretical study of the above, at first, the author aimed at selecting the most relevant material from academic sources such as Google scholar, Elsevier, Scopus as well as other related databases such as the Research gate. Specific criteria were set for the sorting of these papers as part of the quality assessment.

Also, when it comes to the suitability and the quality of each paper, additional criteria were applied. These are the following:

- Question 1: Is the paper properly structured?
- Question 2: Is the paper relevant?
- Question 3: Does the paper come up with conclusions using an appropriate methodology?

To assess the criteria, the following method is used:

For question 1:

- All papers that are used must have an introduction, a main part and a conclusion
- All papers have an abstract
- The full citation is available

For question 2, the following are assessed:

- all papers are recent, meaning that they have been published after 2014
- all papers have more than 20 citations and are reviewed before their publication
- the material used is either academic papers or conference papers and / or official reports that were published by officially recognized institutions (i.e. NATO, the EU, the UN etc).

For question 3:

- The methodology is clear
- The data (if used) is included and / or appropriately cited
- The conclusions are followed by limitations

Then, the keywords for each of the above categories are selected words and terms that, based on a first review of the available academic literature, seem to be fully consistent with the topic. Thus, the following are selected as keywords in order to discuss the HS and its importance, impact and different aspects:

- Homeland security
- Homeland security terrorism
- Homeland security cyberterrorism

Particular emphasis is placed on the last two categories of studies, namely articles and studies on terrorism and cybersecurity in relation to terrorism.

Then, for the second aspect and the second level of analysis, ie the use of knowledge graphs, material was searched with the following keywords:

- knowledge graphs homeland security
- use of knowledge graphs security and defense
- knowledge graphs security terrorism

Based on the material identified in the content analysis, emphasis is placed on the results and conclusions of the literature selected for analysis. Last, it is noted that the material is not selected based on the own assumptions and viewpoint of the author and neither does it reflect the opinions of the members of the university.

2.2 Homeland Security

As it is mentioned above, the main methodology of this chapter is the S.L.R. Overall, the number of available sources were thousands in Google Scholar alone. However, the

following are applicable to the paper and have been later on rejected based on the aforementioned criteria and given that a vast majority only included references to other papers on HS and / or were students' papers and theses and not actual papers on HS. Out of the 59.000 papers and books, 40 were assessed for the paper and 10 are used as they fit all the criteria set above. The material selected for the study of HS is presented in the following table. The papers that have been reviewed yet not included in this S.L.R. are presented in a full table in the appendix of the study.

Table 1: Papers used for the homeland security literature review

Id	Reference (brief)	No. of citations	Type of material	Methodology	Keywords	Topic
1	(Awan & Memon, 2016)	20	Conference paper	Critical review	Homeland security and cyber terrorism	Pakistan and the HS challenges – cyber security in the Middle East
2	(Gardner, 2014)	20	Paper	Case law	Homeland security	Review of law on HS
3	(Gruenewald, Allison-Gruenewald, & Klein, 2015)	30	Paper	Critical literature review	Homeland security terrorism	Assessment of national policies
4	(Haynes & Giblin, 2014)	21	Paper	Systematic data collection (350 small agencies)	Homeland security	Police preparedness

5	(Hiemstra, 2014)	38	Paper	Critical literature review	Homeland security	Critique on HS practices
6	(Kahan, 2015)	27	Paper	Critical literature review	Homeland security	Definition of HS and resilience
7	(Kaunert & Leonard, 2019)	22	Paper	Critical literature review	Homeland security terrorism	Terrorism and HS in the EU
8	(Kelarestaghi, Heaslip, Khalilikhah, Fuentes, & Fessman, 2018)	21	Paper	Critical literature review	Homeland security cyber terrorism	Cyber terrorism and protection of databases
9	(Sageman, 2014)	318	Paper	Critical review	Homeland security terrorism	Critique on the field and methods of terrorism studies
10	(Le & Hoang, 2016)	26	Conference paper	Critical review	Homeland security cyber terrorism	Cyber security metrics

The first article studied for the purposes of this dissertation concerns the study of Hiestra (2014) in relation to the application of HS for immigration management and detention system in the United States. Specifically, the researcher argues that the detention system is, nowadays, part of the government's imaginary and apparatus to achieve HS. More specifically, Hiestra (2014) considers that the implementation of such practices

encourages the development of xenophobic behaviors and exacerbates internal problems in the United States. To support this argument, he critically interprets the concept of HS. Thus, they argue that this term is used as a synonym for “national security” and, therefore, in public discourse and political rhetoric, is associated with the fight against terrorism and crime. As a result, for the researcher, it is considered that policies to ensure internal stability in the name of HS tend to project a specific, patriarchal model, which connects America with a) whites, b) heterogeneity, c) nuclear family and the ideals associated with this particular notion of “security” that precludes the different (Hiemstra, 2014).

Next, one can focus on the study of Kahan (2015) which investigates how and why the US invests in HS, especially from September 11, 2011 onwards. This article explains that, in the context of the implementation of policies to enhance security and resilience in the United States, different and related policies are designed, yet, the officials seem to address various priorities. Simultaneously, it is discussed that there are numerous terms used to describe the concept of HS which, mostly, relate to resilience as well as the concept of resilience in relation to strategy and tactics. The overall definitions gathered by Kahan (2015) are directly related to the subject of the present study and serve to better understand how research and implementation of policies are done in relation to maintaining a high level of security and functionality of institutions. The different interpretations given to resilience therefore concern: the resilience of systems and structures, the stability, the duration, the ability of the state to deal with terrorist threats, the dynamic capabilities and the readiness.

This article further discusses the importance of resilience within structures. Therefore, according to Kahan (2015), HS analysis concerns the a) individuals, b) infrastructure, c) actors, d) systems and e) community. So, in order for there to be duration, stability and efficiency, there must also be communication, consistency of policies and setting clear goals. Accordingly, for the strategy and organization of the institutions it is necessary to implement policies for the development of resilience of different institutions and institutions in order to maximize security and stability. Last, it is essential to plan to be able to respond immediately to crises and emergencies, as well as measure performance with specific indicators, standards and criteria (Kahan, 2015).

On the other hand, Sageman (2014) seems to focus more on the ethical dilemmas posed above for science and the media and less for the general public. More specifically, they point out that terrorism studies and research are mistakenly thought to have started in the United States in 2001, which he believes leads to the notion that terrorism is linked to the teaching of the Qur'an Islam in general. As a consequence, as discussed above, in Kahan (2015)'s article HS seems to be equated with youth mobilization and the treatment or "rescue" of Muslims in the US and outside the US (Britain, Middle East, EU). In addition, the author addresses the fact that the structures and agencies created after 2001, the statements of officials as well as the attitude of the media shape the consciousness of citizens and influence research and scientific development in the field of strategic and political studies, therefore, special attention is required from researchers.

Next, the importance of the evolution of law and legislation in relation to homeland security in the US is studied. Gardner's (2014) research, on which this analysis is based, shows that the evolution of case law in this area is rapid, especially in the period 2001-2008, which is attributed, as Kahan (2015) analyzes, to the increase of external and internal threats to the US. In detail, Gardner (2014) notes that the first appearance of the term seems to be attributed to the letter of President Bush in 2002 who chose the term "homeland" and "national security" instead of "federal", which translates into an attempt to Recognition of the "shared responsibility" of all American citizens for security threats in the country. During the Bush administration, the DHS (Department of Homeland Security) was created, which, according to the researcher, aims to promote cooperation at the federal, state and local levels in addressing the threat of terrorism and immigration. Flows. However, from 2008 onwards, the power and obligation of the police to intervene in matters relating to illegal immigration was reduced and, thus, the need arose to institutionalize the role of each of the law enforcement agencies in the United States. The Printz ruling is considered a milestone because it was discussed before the Supreme Court whether state power has limits and whether there should be a concentration of power in times of insecurity, as well as issues of recruitment, majority power and interpretation of the constitution.

In detail, for the importance of this decision, the researcher notes the following (Gardner, 2014):

1) national security, homeland security and national emergency are not synonymous terms and must be approached without allowing a hegemonic minority to accumulate power
2) when authorities, states and institutions disagree, consensus must be reached
3) no emergency can lead to the violation of the rights set out in the constitution

Having analyzed both the concept and importance of preparedness (Hiemstra, 2014; Kahan, 2015) and the role of institutions (Gardner, 2014), it is possible to assess at this point whether police authorities are able to manage the threats that the US face and are directly linked to HS. More specifically, Haynes & Giblin (2014) investigate how police authorities, especially in small police departments, prepare to manage risks mainly related to terrorism and organized crime. For this reason, the researchers collected data from three hundred and fifty small police units and focused on the following questions:

a) what are the external risks and what are the internal risks for which the asymptomatic are prepared?
b) how is HS related to risk management?
c) what are the main dimensions on which these bodies focus theoretically and practically?

Based on this research, it appears that, ultimately, organizations tend to take measures to address asymmetric external threats focusing on homeland security, time management and reducing the vulnerability of institutions. It is also found, however, that, in essence, risk management strategies do not appear to increase safety, preparedness and flexibility, and the risk under consideration is, in each case, different and subjectively assessed. In particular, each unit tends to be organized on the basis of the personal perceptions of police and armed forces personnel and not on the basis of available data and, in fact, for a specific period of time and not as part of an integrated strategy (Haynes & Giblin, 2014).

Examining the readiness of players at the national level, Gruenewald, Allison-Gruenewald, & Klein (2015) agree with Kahan (2015) and Haynes & Giblin (2014) that internal problems in the US worsened after on September 11 and claim that HS was directly involved in dealing with the terrorist threat, as found by the SLR which has been done in this section. The researchers also showed that the principles of Situational Crime Prevention apply to the study of HS and its applications / consequences as the

two are common and the criteria set for the analysis of events and the assessment of the severity of situations can to be applied to counter-terrorism. These are the following:[2]

- exposed
- vital
- iconic
- legitimate
- destructive
- occupied
- near
- easy

The researchers point out that the aim is for the authorities to be able to "think" as terrorists and, therefore, to identify in time the possible targets of parastatal / terrorist groups and organizations, the risks and the mechanisms. Special emphasis is also placed on the image of terrorists, the viability of policies and measures as well as the legitimacy of the authorities' actions. More specifically, for the purposes of these groups, Gruenewald et al. (2015) showed that the more "obvious" the consequences of a crime, the more likely they are to pose a threat to national security in the US, so civilians are at much greater risk (79%) if they are in public buildings or working for major agencies than if they are ordinary citizens and live outside the center (5%). This means that HS can be adapted to reduce the risks to stability and peace within the US as discussed in the Hiemstra (2014) article and to maximize the efficiency and functionality of institutions as widely stated above (Gardner , 2014; Haynes & Giblin, 2014; Hiemstra, 2014; Kahan, 2015).

The perspectives and possibilities of dealing with crises, the threat of terrorism and external / internal cyber threats are also examined in the study of Kelarestaghi et al. (2018). The researchers propose a complex model for risk assessment developed by the National Institute of Standard and Technology (NIST) – NIST SP 800-30 which is based on three objectives (Kelarestaghi, Heaslip, Khalilikhah, Fuentes, & Fessman, 2018):

[2] The acronym formed by the eight (8) key points of the Situational Crime Prevention approach is "EVIL DONE" (Gruenewald, Allison-Gruenewald, & Klein, 2015, p. 433).

- identifying security challenges and “security gaps” in the system
- identifying threats
- seeking the impact of the hacking incident.

This way, the network administrators will be able to minimize threat and vulnerability and achieve the overall goals of the agencies that can be impacted by cyber security threats. In addition, in their paper, the researchers point out the need to use advanced communication tools and to invest in the creation of networks to reduce both digital / online / cyber threat and physical threats (Kelarestaghi, Heaslip, Khalilikhah, Fuentes, & Fessman, 2018).

The study by Kaunert & Leonard, (2019), which examines how counter-terrorism threats affect the EU and how it responds to its internal threats by applying similar systems to the US, differs significantly from the above articles. In detail, researchers compare the creation of agencies such as DHS (Gardner, 2014) with the strengthening of Europol’s role. Thus, this research shows that, as in the case of the USA, in the EU, bodies and means have been created to deal with terrorism and the common threat, although there is no reference to “homeland” security but EU security and but follow the example of the Bush presidency. Therefore, the EU institutions and, consequently, member states voted in favor of (Kaunert & Leonard, 2019):

- the Directive 2017/541 on combating terrorism in the EU
- agreements for co-operation with the ASEAN
- agreements for co-operation with the states in the Schengen zone and the European Economic Area to address the common challenges faced by the EU and its neighbouring countries.

Similarly, Awan & Memon (2016) research focuses on an environment and subsystem that is not directly relevant to the United States but more indirectly – Pakistan and the Middle East. This research also begins with the assumption that counter-terrorism studies and practices have intensified since 2001, and that, for the United States and its international allies (inside and outside NATO), Emphasis seems to be given to areas such as:

- countering cyber – terrorism
- e-government and the digitization of data and services
- ensuring stability in the international system

- decreasing nuclear threat
- strengthening the position of Middle Eastern countries in the international economy.

Finally, the research of Le and Hoang (2016) examines on a global scale how all the above issues discussed are related. For this purpose, they analyze and evaluate the degree of suitability and application of the Security Maturity Model in information technology and in relation to cyber terrorism. This model is as follows:

A. Managers consider the extent to which the system is secure based on investment in security and sustainability
B. the analysis is divided into five steps and five sets of criteria set:

1) initial level – assessment of the primary practices, the different levels of analysis and requirements of the system
2) repetition and maturity – assessment of the sets of practices multiple times
3) definition of the problem and aims
4) management – particularly software and systems
5) optimization

Thus, it appears that the different levels can be evaluated with different softwares or methods. Based on this research, one understands that tackling cybersecurity-related problems actually starts, as argued above (Gardner, 2014; Haynes & Giblin, 2014; Kaunert & Leonard, 2019) from the grassroots level. The researchers also argue that the structures, methods and systems used by each state in relation to the management of digital systems, security and the operation of the market, among many, determine which systems can be applied in each case. Therefore, because in the US the whole system is oriented towards the achievement of HS, the Security Maturity Model is a suitable tool. The same is true for the EU where the organization and management of institutions and functions are effective and appropriate for tackling external threats, both real and digital (Le & Hoang, 2016).

Having considered the aforementioned conclusions, this essay will attempt to investigate whether knowledge graphs are as an appropriate tool, how they are used and why.

2.3 The use of knowledge graphs for homeland security purposes

The methodology followed for this, second section of the S.L.R. is similar to that of section 2.3. Overall, in the platforms assessed online, 16.800 papers, books and conference materials were identified. Similarly, after a careful examination of the material and an exclusion of students' papers and irrelevant material, out of those, 29 were reviewed and 9 used. The papers that have been reviewed yet not included in this S.L.R. are presented in a full table in the appendix of the study.

Table 2: Papers used for the use of knowledge graphs for homeland security purposes literature review.

Id	Reference (brief)	No. of citations	Type of material	Methodology	Keywords	Topic
1	(Collarana, et al., 2018)	10[3]	Conference paper	Synthesis of framework using models	knowledge graphs defense terrorism	Minte+ project
2	(Hu, Zhang, Liu, & Wang, 2017)	22	Paper	Critical literature review	use of knowledge graphs security and defense	Quantitative methods to assure network security
3	(Iannacone, et al., 2015)	68	Conference paper	Content analysis	Knowledge graphs homeland security	Application of knowledge graphs
4.	(Jones, Bridges, Huffer, & Goodall, 2015)	38	Conference paper	Model creation	Knowledge graphs homeland security	Explaining the algorithmic process used for data extraction

[3] Although the paper has less than 20 citations, it is included in this list due to its scientific value and the impact of the study on the methodology of this essay.

5	(Kaynar, 2016)	52	Paper	Attack graph modeling and taxonomy	use of knowledge graphs security and defense	Study of attack graph paths
6	(Miehling, Rasouli, & Teneketzis, 2015)	55	Conference paper	Automatic approach model / development and application of model	Knowledge graphs homeland security	Application of algorithmic and mathematical models (Bayesian attack graphs) in network security
7	(Noel, Harley, Tam, Limiero, & Share, 2016)	49	Book chapter	Critical literature review	Knowledge graphs homeland security	Explanation of use of a particular type of KGs-CyGraph
8	(Noy, et al., 2019)	55	Book chapter	Critical literature review	use of knowledge graphs security and defense	Methodology of KG – creation and use and different alternatives
9	(Xia & Gu, 2019)	3[4]	Conference paper	Empirical research / KG design	Knowledge graphs homeland security	Building a Terrorist KG

[4] Although this paper has only 3 citations it is included because a) it is directly related to the topic of the S.L.R., b) it is innovative, c) has scientific value for the issues discussed.

Hu et al. (2017) discuss that, as the challenges become more complex, technologies should be developed and quantitative models and tools used to address internal and external threats to states and organizations. In turn, the researchers suggest the use of:

- STS Spatial - Time Sequence - Based Methods, that are a statistical method to process chaotic time series
- GTM Graph Theory Based Models that include, among others, the weighted planning knowledge graphs, Markov's chains and Markov's prediction models and attack graphs
- GAT Game - Theory Based Methods, which are mathematical methods to predict behaviors.

In this section, the second category of models and quantitative methods will be assessed.

Starting from the paper of Iannacone et al. (2015) this concerns the application of knowledge graphs in various fields, especially cybersecurity. This paper analyzes, in particular, how databases are created, how they can be processed by researchers, the capabilities and limitations of KGs. Thus, in relation to the capabilities and usefulness of KGs, it is mentioned, first of all, the usability, the increased security that they provide as well as the possibility for automation of the processes that they offer. Therefore, these systems can be used in the context of HS policies as they address the problem of vulnerability, are flexible, technologically advanced and offer immediate intervention in the event of threats. At the same time, KGs allow for the location of the user and the IP that are blacklisted are excluded immediately. Also, KGs trace relationships, connections and can help locate a malware. The main advantage of their use is that the data can be reused and remodelled (Iannacone, et al., 2015).

Indeed, based on the paper of Noy et al. (2019) one can note that there are a number of alternative models, software and tools that a researcher may use for data processing and the creation of KGs. The key impact of the use of KGs for Noy et al. (2019) is that they allow the user to make data more structured and functional and that they are practical and cost-effective. Also, they mention that the application of KGs does not only concern HS and defense but a wide range of organizations and business sectors, due to the adaptability and usability of the graphs. For this reason, a number of tech companies have developed their own tools to allow the user to create a KG. The models, size and stages of development of each product are summarized as such (Noy, et al., 2019, p. 5):

- Microsoft uses a single ontology to include types of entities, their relations and attributes. The size of the graphs is two billion entities (primary) and around 55 billion facts and the product is currently used by multiple companies.
- Google uses KGs that include a wide range of entities that are strongly typed. The size of the graphs is one billion entities with over seventy assertions and the product is also actively used.
- Facebook uses graphs that allow for a structured insertion and processing of data and optional indexes. Each graph can include over 50 million entities (primary) and 500 million assertions, with the product been currently used.
- Ebay follows the model of Google, yet the size of the graphs is around 100 million produces and the product is under development (2019).
- IBM uses graphs and models that allow the association of various data (entities and relations) and have various sizes. The product is actively used and offered for clients.

The researchers point out that KGs are, indeed, ideal tools to predict behaviours and events. However, they can be challenging as a) they use advanced AI, especially when one processes big data and, therefore, training and expertise on behalf of the user is required, b) the user must make sure that all the data included are correct, which is often an impossible challenge, c) in case a company wants to sell their product, it is necessary that the client be trained as well.

Having considered the use and advantages of the creation of KGs, one can, then, proceed with an analysis of their application in HS. Noel et al. (2016) in their study explain that the tools available today can indeed be used in a variety of ways to enhance network security, preparedness and the ability of operators to deal with cyber threats. More specifically, the authors suggest the use of Cy Graph which is a system created based on these needs and priorities[5]. This system supports the decision-making process,

[5] The function of the Cy Graph and its architecture can be described as such (Noel, Harley, Tam, Limiero, & Share, 2016, p. 7):

1. Ingest: network infrastructure / security posture / cyberthreats / mission dependencies
2. Transformation
3. Analysis through a graph model: dynamics / layering / grouping / filtering / hierarchies
4. Visualization

increases the readiness of the agencies and identifies vulnerability paths. In addition, in case of real threats, Cy Graph sends alerts and suggests ways to resolve the crisis. Its main advantage is that it works with other software, and can be used by real users or by automated systems (digital / computer) (Noel, Harley, Tam, Limiero, & Share, 2016).

For HS the use of Cy Graph can be a solution to address internal and external threats and detect malware, intruders and threats. Already, according to Noel et al. (2016) US DHS, DISA and NIST seem to collaborate and take advantage of the possibilities provided by the use of KG and Cy Graph. The advantage of Cy Graph KGs seems to be that they are used to create a variety of tools and databases as well as graphs for threat analysis. Additional tools are:

- GOTS (Government off-the-shelf attack graph analyses) i.e. TVA and NetSPA
- COTS (commercial off-the-shelf KGs), i.e. Skybox and RedSeal.

The potential data sources include national and commercial vulnerability databases, cyber mission assessments and other relevant sources. Also, Cy Graph is suitable for the analysis and synthesis of big data and to act as a) a direct tool that is usable by a client, b) a server and c) an intermediary (Noel, Harley, Tam, Limiero, & Share, 2016).

The use of databases that specialize in specialized sectors or that are particularly associated with a particular issue are, is of great importance for the management of the risks associated with terrorism in cyberspace or in the physical space. Xia & Gu (2019) recommend the use of KGs to extract data from the Wikipedia website to better understand terrorist attacks and improve the research methodology of this type of data. More specifically, their methodology is as follows:

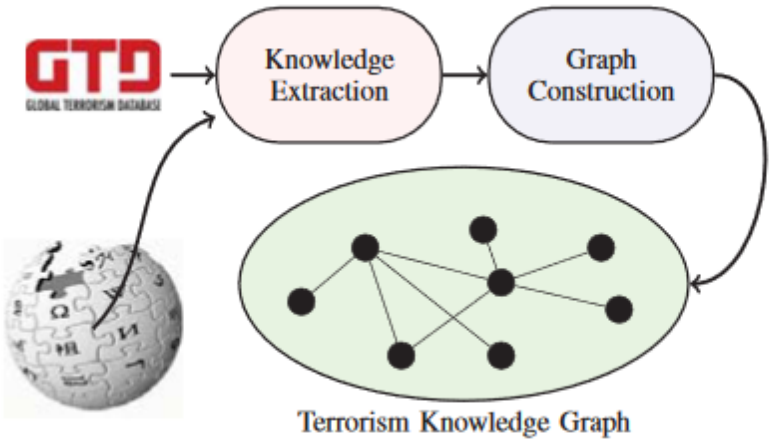


Figure 1: brief description of the process of the creation of the TKG (Xia & Gu, 2019, p. 195), all rights reserved by the authors.

To create the Terrorism Knowledge Graph, Xia & Gu (2019) consider six different parameters which concern:

- The identification and details of the incident
- The "weapon" used for a given threat, attack, incident etc.
- The perpetrator(s), meaning the people or group of people that are responsible for the incident
- The location, therefore the place of the attack (country, region, city etc.)
- The target of the attack
- The damage caused (physical, financial) as well as the total fatalities

Then, the construction of the Terrorist KGs is done using one of three methods: a) extracting data and knowledge from structured data that include country specific or field / hierarchic information, b) free text, that can be used to add notes and extract facts, c) the Wikipedia infobox to extract, pair and organize values. Microsoft excel ™ and other statistics software can be used for the extraction and organization of the data and the different graphs (Xia & Gu, 2019).

Next, Jones, Bridges, Huffer, & Goodall (2015) equally emphasize the fact that data on terrorism, defense and security, in most cases, are not categorized from the outset by the competent authorities or the persons who publish the information. they. Therefore, a more specialized processing of information is required, often from free text, as emphasized by Xia & Gu (2019).

The methods that can be used to resolve this issue are varied. First, Jones et al. (2015) suggest the use of specially designed techniques for automatic search and entry of information, which can be done with relation extraction. This method is a tool for evaluating physical speech to retrieve data related to the name, location and other relevant information about an object. This information covers what is stated in their texts and Iannacone, et al., (2015) and Noel, et al. (2016). Alternatively, each researcher or developer can create a new algorithm to process and categorize data. The key method to achieve that is through relations and patterns. The researcher either enters (manually or with the use of a software) all words that are part of speech or selected words and parts of speech. Afterwards, a parse tree path is created to discover the patterns and databases (Jones, Bridges, Huffer, & Goodall, 2015).

At this point, also, it is interesting to add some notes on the study of Collarana, et al. (2018) who have used mathematical and computational models as part of the MINTE+ framework[6] that can be used in policy making, market research, manufacturing and law among others. The researchers suggest that such models can be used to process various data from different databases that are complex and related to (Collarana, et al., 2018):

- Search for activity of specific individuals (i.e. politicians) online.
- Location of fanatics and terrorists and monitoring of their activities online
- Posts, add and relevant material on illegal trade, selling and distribution of drugs, guns etc.

Next, more focused on creating specialized models for defense and security is the article by Miehling, Rasouli, & Teneketzis (2015) which addresses the need to create effective, independent and flexible models for predicting potential security issues. More specifically, this study focuses on the development of software and protocols for the protection of national systems in the US, which is a key part of the policy and objectives of DHS and its respective teams (ICT - CERT). The researchers point out that perhaps the biggest threat and malfunction of the systems currently in place in the United States is that vulnerabilities are numerous and therefore it is difficult to defend the cyberspace and the networks across the US. Also, the handling of the network at the moment is an additional challenge for the "defenders" due to the fact the fact that a) the network is vulnerable and b) the paths are infinite, therefore, it allows the successful outcome of the perpetrators' attempts to violate security systems (Miehling, Rasouli, & Teneketzis, 2015).

The attack graphs studied by Miehling, Rasouli, & Teneketzis (2015) are graphs that can cover a lot of information but, typically, are large and difficult to create and to be monitored by a single user. It is emphasized that the MTD (Move Target Defense) type schemes can solve many of these issues as they are models used for defense and safety and are dynamically adapted to different situations. In the same lines, Kaynar (2016) article classifies the various attack graphs that can be used and the modeling process to create them. Like Noy et al. (2019) so Kaynar (2016) agrees that the first step in creating attack graphs is the determination of the initial privileges and the goals of the attacker. Then, the users can select among a number of tools and methods to determine

[6] The MINTE+ framework is RDF based and used for the synthesis of data.

which one is more applicable. The conditions, vulnerabilities, historical data as well as the templates and models can enable the user to be more functional and to tackle the threat timely and effectively.

Typically, these models can be utilized to locate multiple intruder locations, reduce the intruder's ability to breach databases, or allow the operator (defender) to temporarily alter network characteristics to ensure network protection. The problem, in this case, is that network connectivity and responsiveness are affected (Miehling, Rasouli, & Teneketzis, 2015).

The researchers, in this case, use Bayesian Attack Graphs which result from the application of mathematical models (probabilities) and statistical analysis. The mathematical investigation of the possibility of attack allows the immediate response of the defender. Essentially, the user (defender) examines based on the BAGs the capabilities of the attacker at a given time (t) and his attributes and then adjusts the reaction of the entire system and network. The attributes are usually numerous although key attributes are identified which are the main focus of the attacker. In a typical model, the defender will have specific options, such as disconnecting the system or blocking a specific move of the attacker. Therefore, a cost estimate is made using a BAG and then the best possible solution is selected.[7]

[7] Note: due to the fact that the model used in the study of Miehling et al. (2015) is not directly applicable to the main methodology of the present study, the full model is not included in this S.L.R.

3 Materials and Methods

In this chapter I will discuss the technology stack used for the implementation of the experiment, as well as the datasets used, along with the data cleaning processes for each dataset.

3.1 Technology

The used technologies are divided into two parts. Initially for the data cleaning and the data analysis I am using Python v.3.8.5 and Jupyter Notebook v.6.03, through Anaconda Navigator, along with some unorthodox, yet effective techniques involving Notepad and Microsoft Excel 2019, in order to address obstacles described later.

For the construction of the knowledge graphs, I am using Neo4j Desktop Community Edition v.1.3.11, Graph Database v.4.2.1, and for the scripting I am using the Neo4j's specific language Cypher. While processing our data we need to keep in mind Neo4j's limitations, only csv data can be imported. Regarding the dataset, I opted to eradicate the null values from the RAND database, while I preserved them in the GTD, this way, two different approaches can be showcased.

3.2 RAND Database of Worldwide Terrorism Incidents

The first dataset used for the purposes of this thesis is the RAND Database of Worldwide Terrorism Incidents.

3.2.1 Description

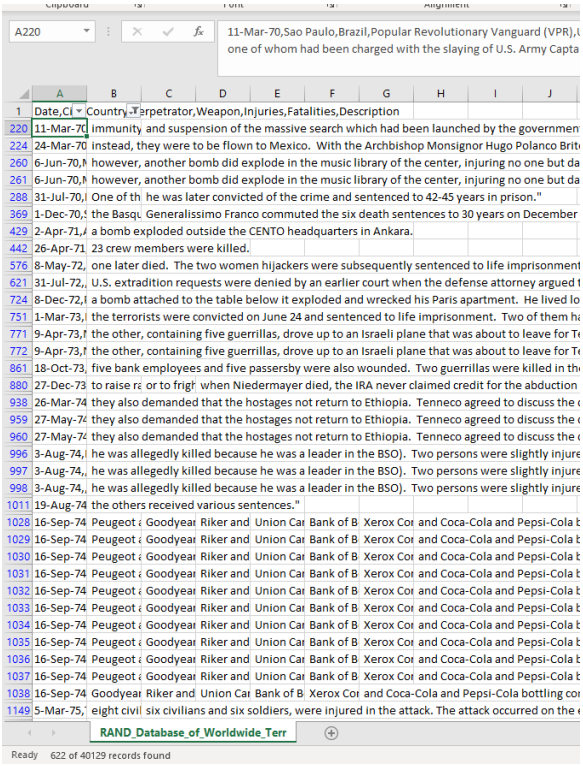
The RAND Database of Worldwide Terrorism Incidents contains data pertaining to Terrorism related incidents from 1968 to 2009 (RAND Corporation, nd.). This dataset contains valuable information on various aspects of terrorism. It contains over 40,000 incidents of terrorist attacks, incorporated from various sources. The RAND Corporation offers the day to pay the database free and publicly in order to help researchers and analysts achieve informed predictions.

The data set contains 8 distinct values:

- Date on which the incident happened,
- City in which the incident happened,
- Country in which the City is,
- Perpetrator of the incident,
- Weapon used,
- Number of Injured,
- Number of Fatalities, and
- Description of the incident.

3.2.2 Data Cleaning

The dataset contains 40129 instances. The first problem that is encountered with the dataset is that, even though it is a csv file, meaning that all the data should be in the first column, separated by commas, it is not the case. If we use the Excel filter function in the second column, excluding the blank cells we get 622 instances that have their Description column spanning in multiple cells.



A220

11-Mar-70,Sao Paulo,Brazil,Popular Revolutionary Vanguard (VPR),
one of whom had been charged with the slaying of U.S. Army Capta

	A	B	C	D	E	F	G	H	I	J
1	Date,Ci	Country erpetrator,Weapon,Injuries,Fatalities,Description								
220	11-Mar-70	immunity	and suspension of the massive search which had been launched by the governmen							
224	24-Mar-70	instead, they were to be flown to Mexico. With the Archbishop Monsignor Hugo Polanco Brit								
260	6-Jun-70,	however, another bomb did explode in the music library of the center, injuring no one but da								
261	6-Jun-70,	however, another bomb did explode in the music library of the center, injuring no one but da								
288	31-Jul-70,	One of th	he was later convicted of the crime and sentenced to 42-45 years in prison."							
369	1-Dec-70,	the Basqu	Generalissimo Franco commuted the six death sentences to 30 years on December							
429	2-Apr-71,	a bomb exploded outside the CENTO headquarters in Ankara.								
442	26-Apr-71	23 crew members were killed.								
576	8-May-72,	one later died. The two women hijackers were subsequently sentenced to life imprisonment								
621	31-Jul-72,,	U.S. extradition requests were denied by an earlier court when the defense attorney argued								
724	8-Dec-72,	a bomb attached to the table below it exploded and wrecked his Paris apartment. He lived lo								
751	1-Mar-73,	the terrorists were convicted on June 24 and sentenced to life imprisonment. Two of them ha								
771	9-Apr-73,	the other, containing five guerrillas, drove up to an Israeli plane that was about to leave for T								
772	9-Apr-73,	the other, containing five guerrillas, drove up to an Israeli plane that was about to leave for T								
861	18-Oct-73,	five bank employees and five passersby were also wounded. Two guerrillas were killed in th								
880	27-Dec-73	to raise r	or to frigh	when Niedermayer died, the IRA never claimed credit for the abduction						
938	26-Mar-74	they also demanded that the hostages not return to Ethiopia. Tenneco agreed to discuss the								
959	27-May-7	they also demanded that the hostages not return to Ethiopia. Tenneco agreed to discuss the								
960	27-May-7	they also demanded that the hostages not return to Ethiopia. Tenneco agreed to discuss the								
996	3-Aug-74,	he was allegedly killed because he was a leader in the BSO). Two persons were slightly injure								
997	3-Aug-74,,	he was allegedly killed because he was a leader in the BSO). Two persons were slightly injure								
998	3-Aug-74,,	he was allegedly killed because he was a leader in the BSO). Two persons were slightly injure								
1011	19-Aug-74	the others received various sentences."								
1028	16-Sep-74	Peugeot	Goodyear	Riker and	Union Car	Bank of B	Xerox Cor	and Coca-Cola and Pepsi-Cola		
1029	16-Sep-74	Peugeot	Goodyear	Riker and	Union Car	Bank of B	Xerox Cor	and Coca-Cola and Pepsi-Cola		
1030	16-Sep-74	Peugeot	Goodyear	Riker and	Union Car	Bank of B	Xerox Cor	and Coca-Cola and Pepsi-Cola		
1031	16-Sep-74	Peugeot	Goodyear	Riker and	Union Car	Bank of B	Xerox Cor	and Coca-Cola and Pepsi-Cola		
1032	16-Sep-74	Peugeot	Goodyear	Riker and	Union Car	Bank of B	Xerox Cor	and Coca-Cola and Pepsi-Cola		
1033	16-Sep-74	Peugeot	Goodyear	Riker and	Union Car	Bank of B	Xerox Cor	and Coca-Cola and Pepsi-Cola		
1034	16-Sep-74	Peugeot	Goodyear	Riker and	Union Car	Bank of B	Xerox Cor	and Coca-Cola and Pepsi-Cola		
1035	16-Sep-74	Peugeot	Goodyear	Riker and	Union Car	Bank of B	Xerox Cor	and Coca-Cola and Pepsi-Cola		
1036	16-Sep-74	Peugeot	Goodyear	Riker and	Union Car	Bank of B	Xerox Cor	and Coca-Cola and Pepsi-Cola		
1037	16-Sep-74	Peugeot	Goodyear	Riker and	Union Car	Bank of B	Xerox Cor	and Coca-Cola and Pepsi-Cola		
1038	16-Sep-74	Goodyear	Riker and	Union Car	Bank of B	Xerox Cor	and Coca-Cola and Pepsi-Cola bottling co			
1149	5-Mar-75,	eight civil	six civilians and six soldiers, were injured in the attack. The attack occurred on the							

622 of 40129 records found

Figure 2: RDWTI Instances where the last column spans in multiple cells.

In order to address this issue, an unorthodox, yet simple and effective approach was taken. We copied the entirety of the dataset in a txt file. The Notepad application automatically inserts a Tab space when there are multiple columns. This way a simple Replace function, where there is a Tab space with a single space merges all the lines that have multiple columns, while at the same time leaving intact the lines that are properly formatted. Finally, we save the txt file, for good measure, while we copy the entire dataset back to a new csv file.

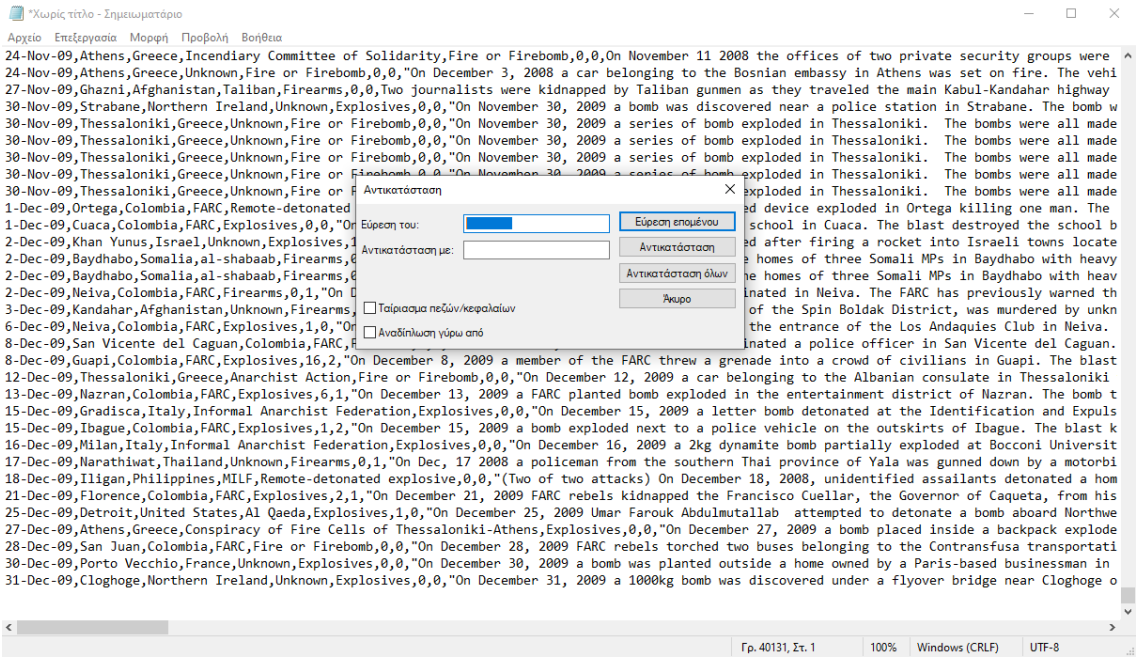


Figure 3: Notepad Replace function.

The next problem we face, is when we load the RDWTI on a Pandas DataFrame in Python. As we see in Figure 4, in lines 3, 4, 5, 6, 7 and 8, out of the first 10 lines, all the information is placed in the first column, while leaving all the other columns with null values. After experimentation, I have concluded that, this happens because those lines have cells within quotation marks, because their value contains one, or multiple commas, e.g. line 4, City “Washington, D.C.” or line 3, Description “CHILE. An explosion from a single stick of dynamite went off on the patio of the Santiago Binational Center, causing $21,000 in damages”.

```
In [16]: df = pd.read_csv("C:/Users/emman/Desktop/RDWTI/RDWTI1.csv")
         df.head(10)
```

Out[16]:

	Date	City	Country	Perpetrator	Weapon	Injuries	Fatalities	Description
0	9-Feb-68	Buenos Aires	Argentina	Unknown	Firearms	0.0	0.0	ARGENTINA. The second floor of the U.S. embas...
1	12-Feb-68	Santo Domingo	Dominican Republic	Unknown	Explosives	0.0	0.0	DOMINICAN REPUBLIC. A homemade bomb was found...
2	13-Feb-68	Montevideo	Uruguay	Unknown	Fire or Firebomb	0.0	0.0	URUGUAY. A Molotov cocktail was thrown outsid...
3	20-Feb-68,Santiago,Chile,Unknown,Explosives,0,...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	21-Feb-68,"Washington, D.C.",United States,Unk...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	21-Feb-68,Neot Hakikar,Israel,Unknown,Unknown,...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	22-Feb-68,Quito,Ecuador,Unknown,Explosives,0,0...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	24-Feb-68,Masada,Israel,Other,Explosives,0,0,"...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	5-Mar-68,Riohacha,Colombia,National Liberation...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	8-Mar-68	Rosario	Argentina	Frente de Liberacion Nacional del Vietnam del Sur	Firearms	0.0	0.0	ARGENTINA. The USIS office in Rosario was mac...

Figure 4: Problem when loading the RDWTI on a Pandas DataFrame.

In order to solve this issue, I exported the DataFrame into an xlsx document. There using the replace function, I removed all the commas (68471 replacements), and subsequently all the double (52338 replacements) and single (11990) quotation marks. Due to the elimination of all the commas, and as a result, the quotation marks, the data can be properly loaded.

Before the data is loaded, I used the Excel Filter function, in order to manually add the values of missing cells, as Unknown. There were 4977 instances where the City was missing, 4 instances where the Perpetrator was missing, 3 instances where the Weapon was missing, and 2 instances where the Description was missing.

```
In [10]: df = pd.read_csv("C:/Users/emman/Desktop/RDWTI/RDWTI2.csv", sep=';')
         df.head(10)
```

Out[10]:

	ID	Date	City	Country	Perpetrator	Weapon	Injuries	Fatalities	Description
0	0	9-Feb-68	Buenos Aires	Argentina	Unknown	Firearms	0	0	ARGENTINA. The second floor of the U.S. embas...
1	1	12-Feb-68	Santo Domingo	Dominican Republic	Unknown	Explosives	0	0	DOMINICAN REPUBLIC. A homemade bomb was found...
2	2	13-Feb-68	Montevideo	Uruguay	Unknown	Fire or Firebomb	0	0	URUGUAY. A Molotov cocktail was thrown outsid...
3	3	20-Feb-68	Santiago	Chile	Unknown	Explosives	0	0	CHILE. An explosion from a single stick of dy...
4	4	21-Feb-68	Washington D.C.	United States	Unknown	Explosives	0	0	UNITED STATES. The Soviet embassy was bombed ...
5	5	21-Feb-68	Neot Hakikar	Israel	Unknown	Unknown	0	0	ISRAEL. Palestinian terrorists damaged a pipe...
6	6	22-Feb-68	Quito	Ecuador	Unknown	Explosives	0	0	ECUADOR. A bomb exploded in the Quito Binatio...
7	7	24-Feb-68	Masada	Israel	Other	Explosives	0	0	ISRAEL. Palestinian terrorists fired five mor...
8	8	5-Mar-68	Riohacha	Colombia	National Liberation Army of Colombia (ELN)	Unknown	0	0	COLOMBIA. Three members of the Ejercito de Li...
9	9	8-Mar-68	Rosario	Argentina	Frente de Liberacion Nacional del Vietnam del Sur	Firearms	0	0	ARGENTINA. The USIS office in Rosario was mac...

Figure 5: The final version of the DataFrame

After all this procedure, the dataset is ready to be used.

3.2.3 Ontology

In order to construct the graph, an ontology needed to be created beforehand. This is the design I opted for.

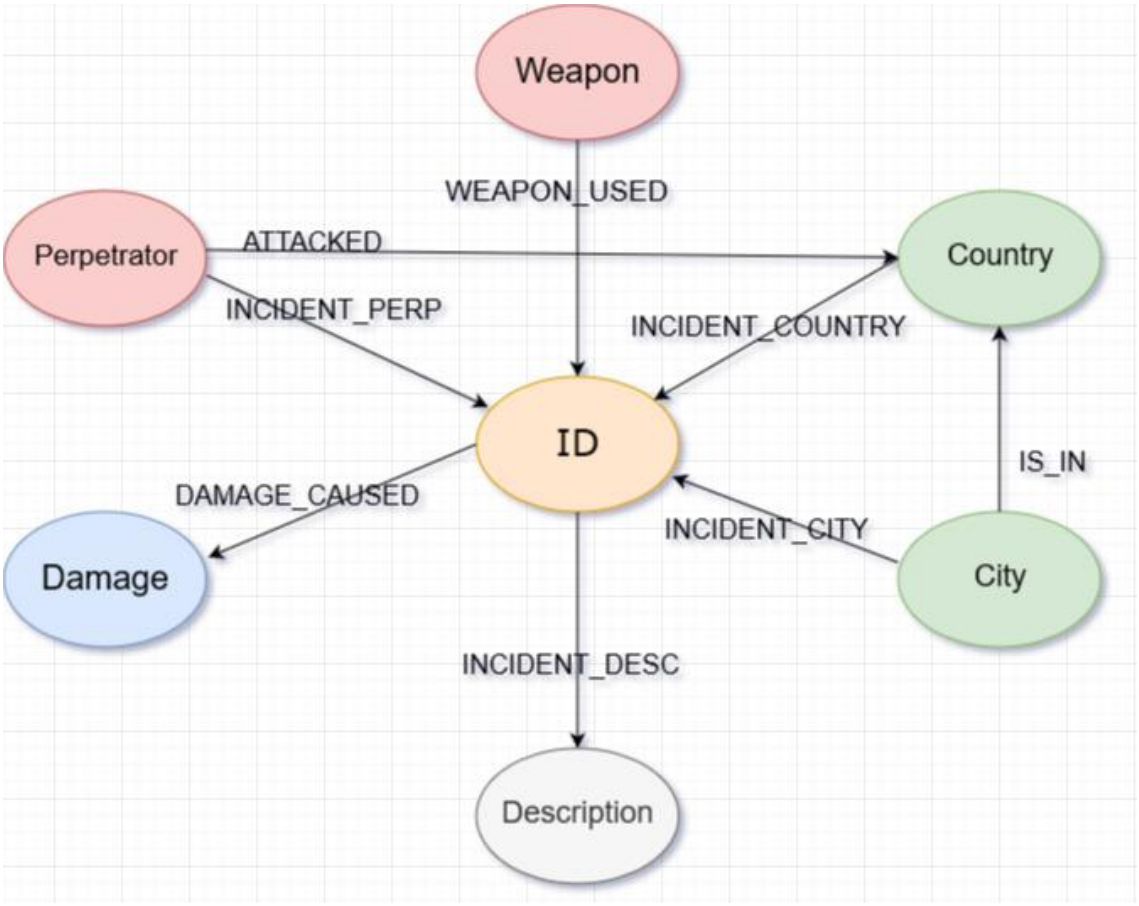


Figure 6: Ontology for RAND Database of Worldwide Terrorism Incidents.

3.3 Global Terrorism Database

The second dataset used for the purposes of this thesis is the Global Terrorism Dataset.

3.3.1 Description

The Global Terrorism Database contains data pertaining to Terrorism related incidents from 1970 to 2018 (Lafree et al., 2006; LaFree et al., 2014; University of Maryland., 2019). This dataset contains valuable information on various aspects of terrorism. It contains over 191,000 incidents of terrorist attacks, along with 135 attributes, incorporated from various sources.

The information in GTD can be summarized and divided into nine general areas:

- GTD ID and Date
- Incident Information
- Incident Location
- Attack Information
- Weapon Information
- Target/Victim Information
- Perpetrator Information
- Casualties and Consequences
- Additional Information and Sources

Going into more detail Table 3 includes all the variables in the Global Terrorism Database, along with their Explanation.

Table 3: Name of Variables, with their Explanation, in GTD (University of Maryland., 2019)

NAME	**Explanation**
eventid	The ID of the incident
iyear	The year of the incident
imonth	The month of the incident
iday	The day of the incident
approxdate	The approximate date of the incident, if the actual date is not recorded
extended	If the incident lasted more than 24 hours
resolution	The date that the incident was resolved, if it was extended
country	Country Code
country_txt	The name of the country
region	Region Code
region_txt	The name of the region
provstate	The name of the first order subnation administrative region
city	The name of the city where the incident took place
latitude	The latitude of the city (WGS1984 Standards)
longitude	The longtitude of the city (WGS1984 Standards)
specificity	Geospatial resolution of latitude and longtitude
vicinity	If the incident occurred in the city itself or the immediate vincinity
location	Additional information about the location
summary	Incident Summary
crit1	First Inclusion Criteria
crit2	Second Inclusion Criteria
crit3	Third Inclusion Criteria
doubtterr	Uncertainty if the Incident should be included

alternative	Alternative Designation Code
alternative_txt	Alternative Designation
multiple	Part of Multiple Incident
success	If the attack was successful, based on its classification
suicide	If the perpetrator did not inded to survive the attack
attacktype1	Method of attack's code
attacktype1_txt	Method of attack's explanation
attacktype2	Second method of attack's code
attacktype2_txt	Second method of attack's explanation
attacktype3	Third method of attack's code
attacktype3_txt	Third method of attack's explanation
targtype1	Type of primary target code
targtype1_txt	Type of primary target
targsubtype1	Specific type of primary target code
targsubtype1_txt	Specific type of primary target
corp1	Name of primary Company or Government Agency targeted
target1	Specific primary person, building, installation etc targeted
natlty1	Nationality of the primary Target code
natlty1_txt	Nationality of the primary Target
targtype2	Type of second target code
targtype2_txt	Type of second target
targsubtype2	Specific type of second target code
targsubtype2_txt	Specific type of second target
corp2	Name of second Company or Government Agency targeted
target2	Specific second person, building, installation etc targeted
natlty2	Nationality of the second Target code
natlty2_txt	Nationality of the second Target
targtype3	Type of third target code
targtype3_txt	Type of third target
targsubtype3	Specific type of third target code
targsubtype3_txt	Specific type of third target
corp3	Name of third Company or Government Agency targeted
target3	Specific third person, building, installation etc targeted
natlty3	Nationality of the third Target code
natlty3_txt	Nationality of the third Target
gname	Perpetrator name
gsubname	Additional details to perpertrator's name
gname2	Second perpetrator name
gsubname2	Additional details to second perpertrator's name
gname3	Third perpetrator name
gsubname3	Additional details to third perpertrator's name
motive	Motive behind the attack
guncertain1	If the involvement of the primary perpetrator is reported by sources

guncertain2	If the involvement of the second perpetrator is reported by sources
guncertain3	If the involvement of the third perpetrator is reported by sources
individual	If the attack was carried out by individuals without affiliations
nperps	Number of Perpetrators
nperpcap	Number of Perpetrators Captured
claimed	If there has been Claim of Responsibility
claimmode	Claim of Responsibility Code
claimmode_txt	Claim of Responsibility Text
claim2	Second Claim
claimmode2	Second Claim of Responsibility Code
claimmode2_txt	Second Claim of Responsibility Text
claim3	Third Claim
claimmode3	Third Claim of Responsibility Code
claimmode3_txt	Third Claim of Responsibility Text
compclaim	More than one groups claiming responsibility
weaptype1	Weapon of attack code
weaptype1_txt	Weapon of attack
weapsubtype1	Sub-type of primary weapon code
weapsubtype1_txt	Sub-type of primary weapon
weaptype2	Second weapon of attack code
weaptype2_txt	Second weapon of attack
weapsubtype2	Sub-type of second weapon code
weapsubtype2_txt	Sub-type of second weapon code
weaptype3	Third weapon of attack code
weaptype3_txt	Third weapon of attack
weapsubtype3	Sub-type of third weapon code
weapsubtype3_txt	Sub-type of third weapon
weaptype4	Fourth weapon of attack code
weaptype4_txt	Fourth weapon of attack
weapsubtype4	Sub-type of fourth weapon code
weapsubtype4_txt	Sub-type of forth weapon
weapdetail	Any extra information regarding the weapons used
nkill	Total number of Fatalities
nkillus	Number of US Fatalities
nkillter	Number of Perpetrator Fatalities
nwound	Total number of Injuries
nwoundus	Number of US Injuries
nwoundte	Number of Perpetrator Injuries
property	Property damage
propextent	Extent of Property Damage
propextent_txt	Category of Extent of Property Damage
propvalue	Value of Property Damage
propcomment	Property Damage Comments

ishostkid	If the victims were taken hostage or kidnapped
nhostkid	Total Number of Hostages/Kidnapping Victims
nhostkidus	Number of US Hostages/Kidnapping Victims
nhours	Hours of Kidnapping/Hostage Incidenty
ndays	Days of Kidnapping/Hostage Incidenty
divert	Country That Kidnappers/Hijackers Diverted To
kidhijcountry	Country of Kidnapping/Hijacking Resolution
ransom	Ransom Demanded
ransomamt	Total Ransom Amount Demanded
ransomamtus	Ransom Amount Demanded from U.S. Sources
ransompaid	Total Ransom Amount Paid
ransompaidus	Ransom Amount Paid By U.S. Sources
ransomnote	Ransom Notes
hostkidoutcome	Kidnapping/Hostage Outcome Code
hostkidoutcome_txt	Kidnapping/Hostage Outcome
nreleased	Number Released/Escaped/Rescued
addnotes	Additional Notes
scite1	First Source Citation
scite2	Second Source Citation
scite3	Third Source Citation
dbsource	Data Collection
INT_LOG	Comparison between nationality of perpetrator and location of attack, logistically driven
INT_IDEO	Comparison between nationality of perpetrator and nationality of target, ideologically driven
INT_MISC	Comparison between nationality of perpetrator and nationality of target, not necessarily ideologically or logistically driven
INT_ANY	Any of the above
related	Related incidents

3.3.2 Data Cleaning

The data contained in the dataset was formatted properly, thus there was no need to proceed with data cleaning. Another important factor contributing to this, is the choice to leave empty cells without filling them.

However, for performance reasons, I opted for slicing parts of the database, in order to load them on Neo4j. The following .csv files have been created:

- City.csv, containing city, country_txt, provstate, region_txt,
- Country.csv, containing country_txt, regiont_txt,
- Perpetrator.csv, containing eventid, gname, motive,

- Provstate.csv, containing provstate, country_txt, region_txt,
- Region.csv, containg region_txt, and
- Target.csv, containing targtype1, targtype1_txt, targsubtype1, targsubtype1_txt, corp1, target1, natlty1, natlty1_txt.

3.3.3 Ontology

In order to construct the graph, an ontology needed to be created beforehand. This is the design I opted for.

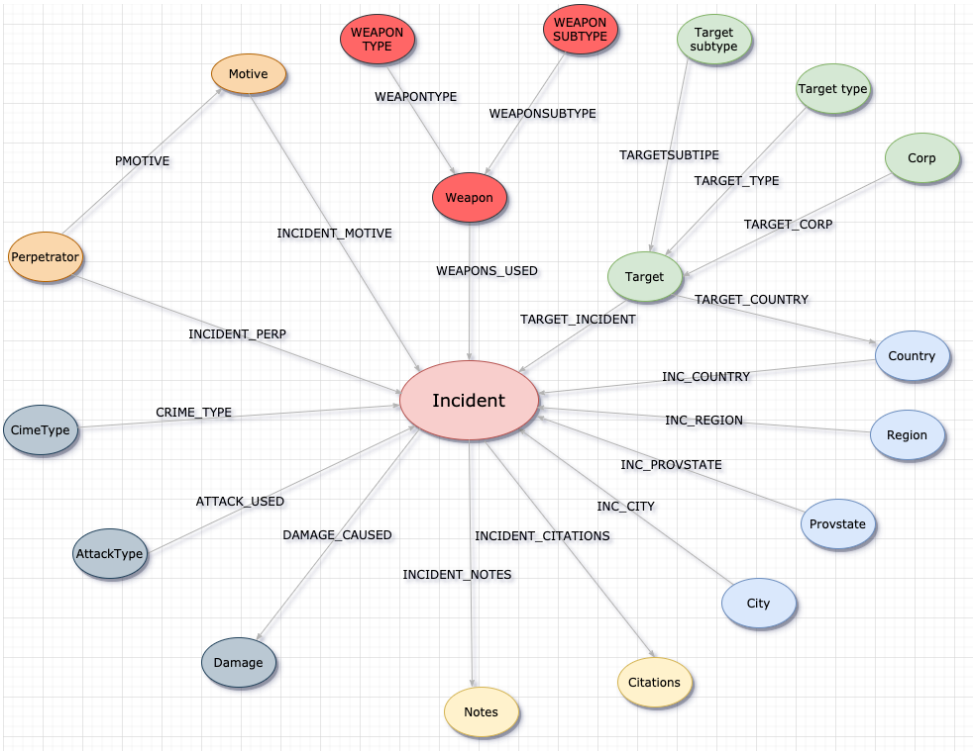


Figure 7: Global Terrorism Database Ontology

4 Results

In this section I will run some indicative queries, to highlight the information that can be extracted by the knowledge graphs.

4.1 RAND Database Graph

1. Get the countries that were attacked by the Taliban

Query:

```
MATCH (n:Perp_Name)-[:ATTACKED]-
(Target_Country) WHERE n.name = "Taliban" RETURN n, Target_Country
```

Figure 8: RDWTI Countries attacked by the Taliban

2. Get all the Perpetrators that attacked Iraq

Query

```
MATCH (Perp_Name)-[:ATTACKED]-
(n:Target_Country) WHERE n.name = "Iraq" RETURN n, Perp_Name
```

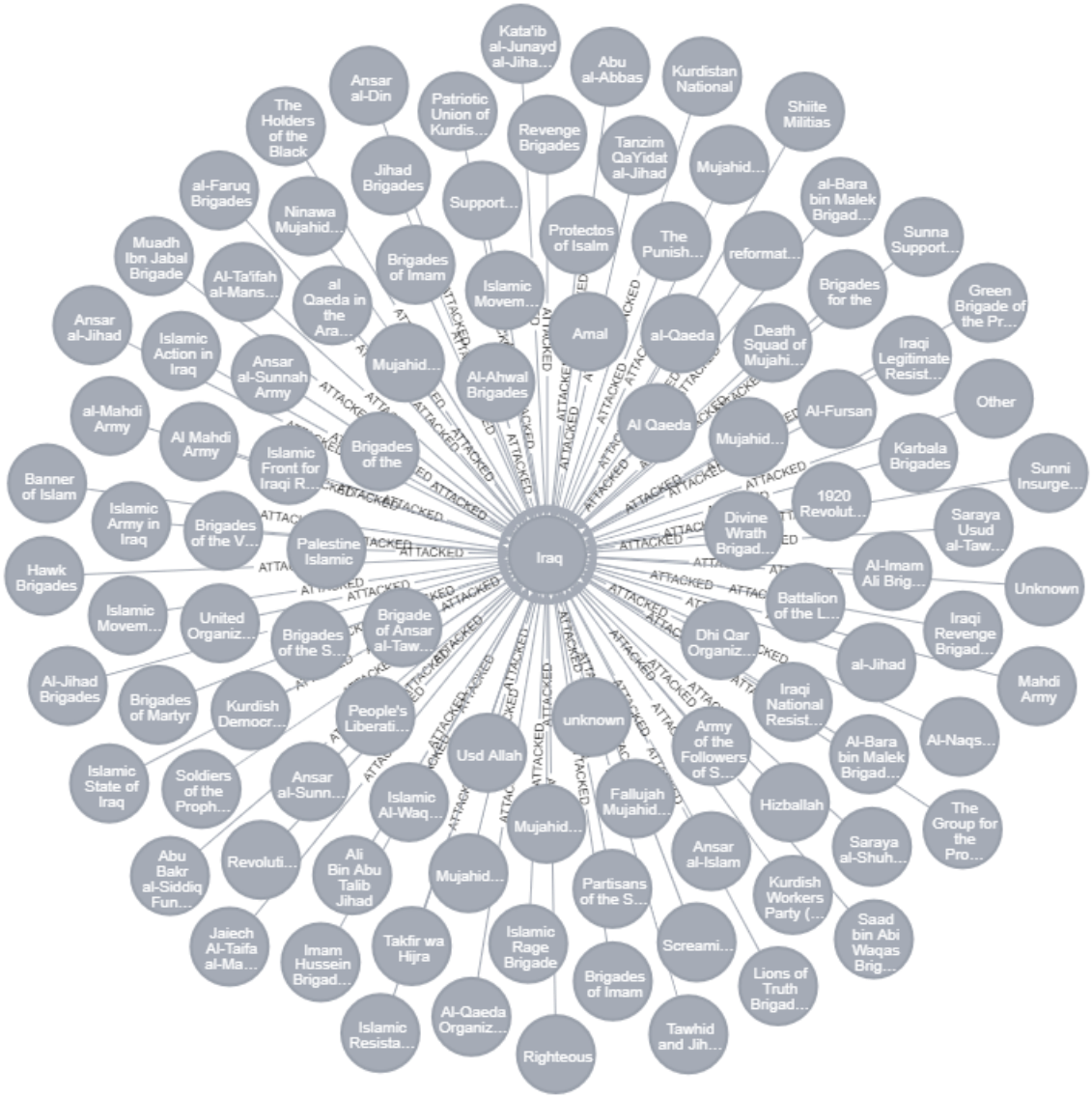


Figure 9: RDWTI Perpetrators that attacked Iraq

3. All the cities in Iraq that had an attack

Query:

```
MATCH (Target_City)-[:IS_IN]-
>(n:Target_Country) WHERE n.name="Iraq" RETURN n, Target_City
```

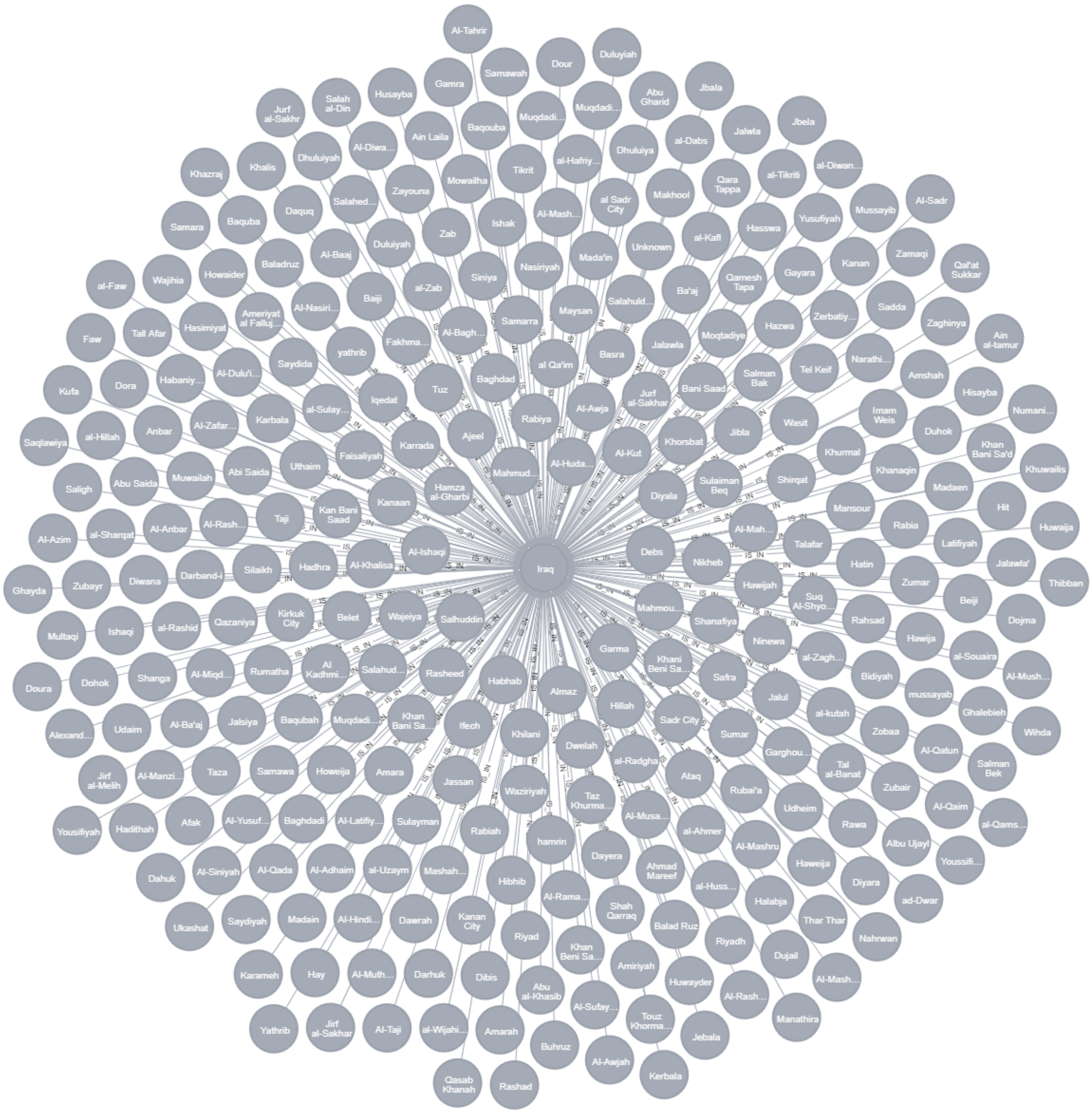


Figure 10: RDWTI All the cities in Iraq that had an attack

4.2 Global Terrorism Database Graph

1. Get the Graph Schema

Query:

CALL db.schema.visualization

Figure 11: GTD Graph Shema

2. Describe a single incident by event id.

Query:

```
MATCH (p:Perpetrator)-[]-(m:Motive),
(p)-[]-(i:Incident)-[]-(w:Weapon)-[]-(wt:WeaponType),
(w)-[]-(ws:WeaponSubtype),
(i)-[]-(t:Target)-[]-(tt:TargetType),
(t)-[]-(ts:TargetSubtype),
(t)-[]-(corp:Corp1),
(i)-[]-(c:City),
(i)-[]-(pr:Provstate),
(i)-[]-(r:Region),
(i)-[]-(cntr:Country),
(i)-[]-(n:Notes),
(i)-[]-(ctn:Citation),
(i)-[]-(d:Damage),
(i)-[]-(at:Attacktype),
(i)-[]-(ct:CrimeType),
(i)-[]-(m)
WHERE i.eventid =200501200002.0
RETURN i,p,m,w,wt,ws,t,tt,ts,corp,c,pr,r,cntr,n,ctn,d,at,ct
```

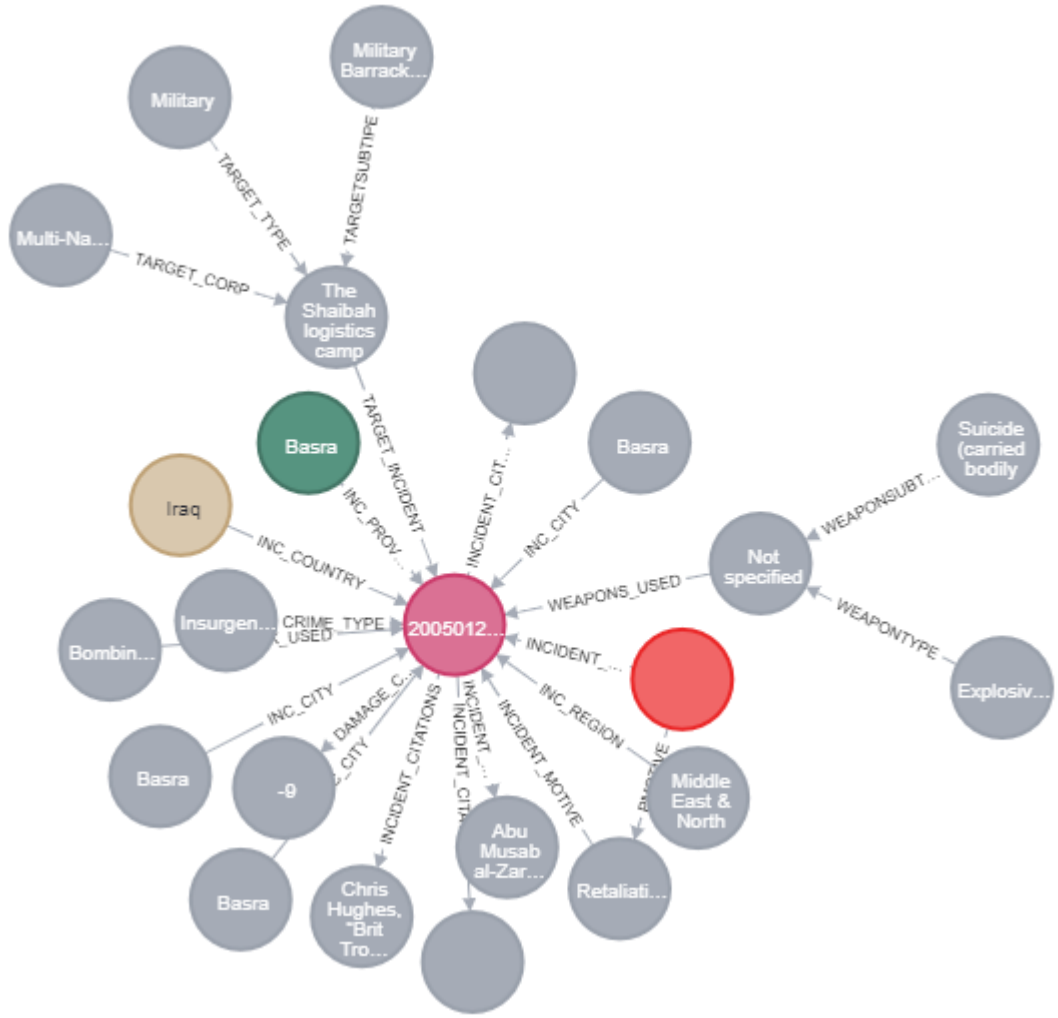


Figure 12: GTD Single incident by event id.

3. All the terrorist groups that attacked Iraq

Query:

```
MATCH q=(p:Perpetrator)-[:INCIDENT_PERP]-(i:Incident)-[:INC_COUNTRY]-
(c:Country)
WHERE c.name="Iraq" and p.description<>"Unknown"
RETURN DISTINCT p.description
```

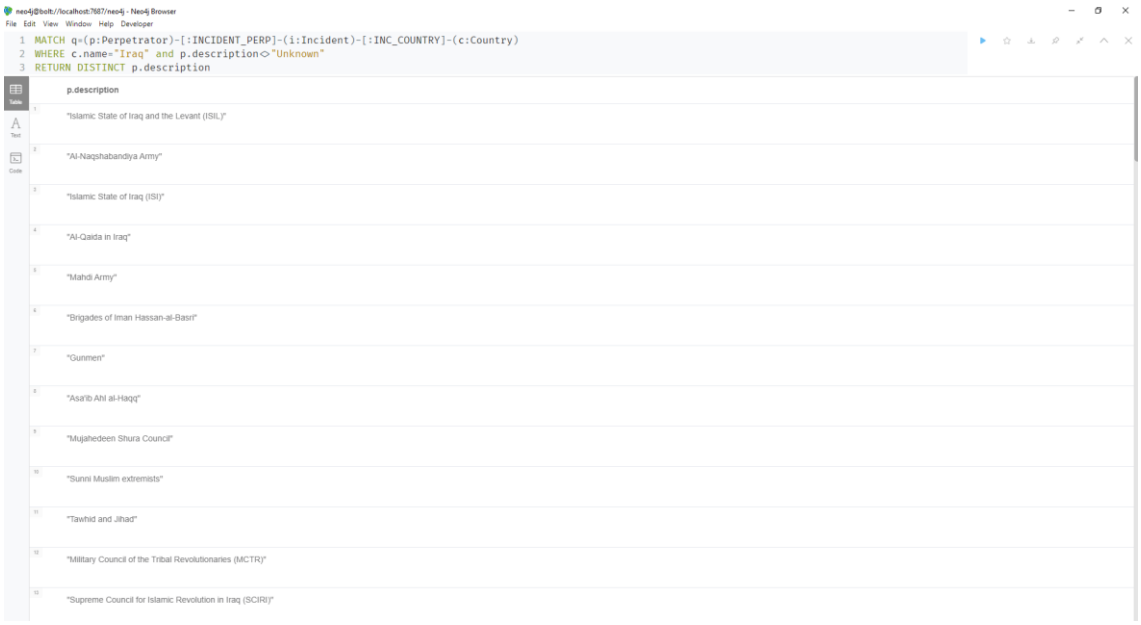


Figure 13: GTD All not unknown terrorist groups that attacked Iraq

4. Attacks in Iraq using explosives and their total number

Query:

```
MATCH q=(wp:WeaponType)-[:WEAPONTYPE]->(w:Weapon)-
[:WEAPONS_USED]->(i:Incident)-[:INC_COUNTRY]-(c:Country)
WHERE c.name="Iraq" AND wp.description= "Explosives"
RETURN i.eventid;
```

```
MATCH q=(wp:WeaponType)-[:WEAPONTYPE]->(w:Weapon)-
[:WEAPONS_USED]->(i:Incident)-[:INC_COUNTRY]-(c:Country)
WHERE c.name="Iraq" AND wp.description= "Explosives"
RETURN count(i);
```

	i.eventid
1	201010110006.0
2	201511290006.0
3	201802170001.0
4	201501160039.0
5	201105300015.0
6	201706300044.0
7	200805120008.0
8	201801310021.0
9	201601280033.0
10	201612010021.0
11	201506100001.0
12	201501030036.0

Figure 14: GTD Attacks in Iraq using explosives

	count(i)
1	19876

Figure 15: GTD Total number of attacks in Iraq using explosives

5. The attacks in Iraq by Al-Qaeda using explosives

Query:

```
MATCH q=(p:Perpetrator)-[:INCIDENT_PERP]-(i:Incident)-[:INC_COUNTRY]-
(c:Country),
(wp:WeaponType)-[:WEAPONTYPE]->(w:Weapon)-[:WEAPONS_USED]->(i)
WHERE   p.description =~'(?i).*Al-
Qaida.*' AND c.name = "Iraq" AND wp.description="Explosives"
RETURN  toInteger(i.eventid);
```

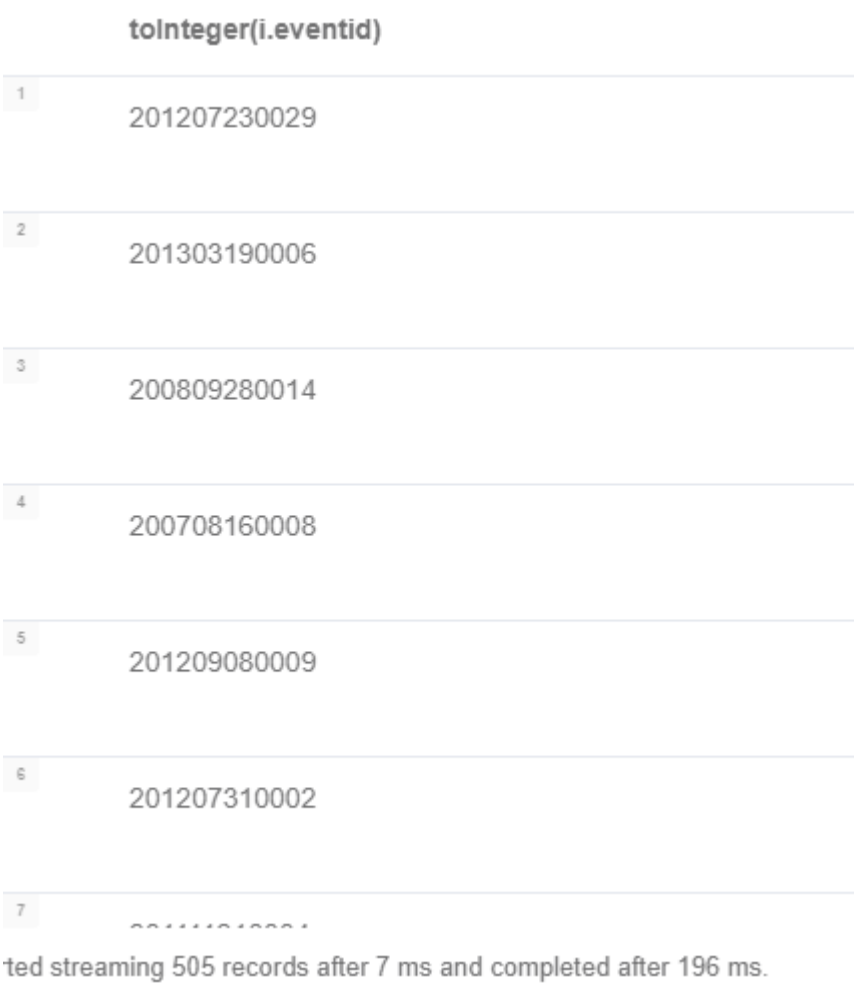


Figure 16: GTD Attacks in Iraq by Al-Qaeda using explosives

6. All the cities Al-Qaeda attacked.

Query:

```
MATCH (p:Perpetrator)-[:INCIDENT_PERP]-(i:Incident)<-[:INC_CITY]-(c:City)
WHERE   p.description =~'(?i).*Al-Qaida.*'
RETURN   DISTINCT c.name
ORDER BY c.name;
```

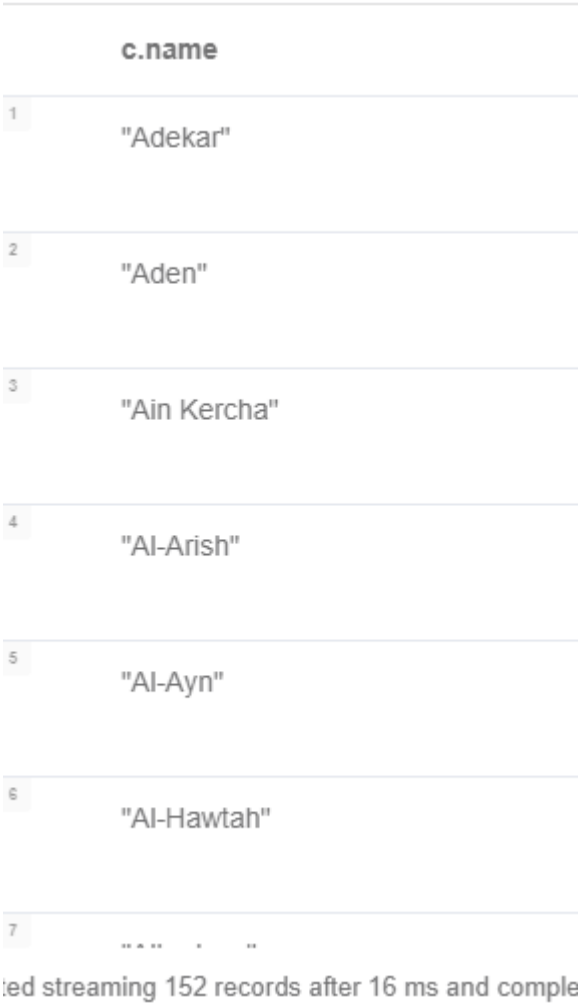


Figure 17: GTD All the cities Al-Qaeda attacked

7. All the incidents where Al-Qaeda was the perpetrator

Query:

```
MATCH q=(p:Perpetrator)-[:INCIDENT_PERP]-(i:Incident)
WHERE   p.description =~'(?i).*Al-Qaida.*'
RETURN  toInteger(i.eventid), p.description;
```

	toInteger(i.eventid)	p.description
1	200109110005	"Al-Qaida"
2	200403110001	"Al-Qaida"
3	200403110003	"Al-Qaida"
4	200205080002	"Al-Qaida"
5	200507210002	"Al-Qaida"
6	200712270001	"Al-Qaida"
7		

rted streaming 2072 records after 15 ms and completed after 20 ms, displaying first 1000 rows.

Figure 18: GTD All the incidents where Al-Qaeda was the perpetrator

8. All the information (nodes) conneceted in some way to Al-Qaeda (Al-Qaeda attacked Iraq, using Explosive, with Damage 5, and Crimetype Terror attack), limited to the first 100 incidents, due to computational stress.

Query:

```
MATCH (p:Perpetrator)-[]-(m:Motive),
(p)-[]-(i:Incident)-[]-(w:Weapon)-[]-(wt:WeaponType),
(w)-[]-(ws:WeaponSubtype),
(i)-[]-(t:Target)-[]-(tt:TargetType),
(t)-[]-(ts:TargetSubtype),
(t)-[]-(corp:Corp1),
(i)-[]-(c:City),
(i)-[]-(pr:Provstate),
(i)-[]-(r:Region),
(i)-[]-(cntr:Country),
(i)-[]-(n:Notes),
(i)-[]-(ctn:Citation),
```

```
(i)-[]-(d:Damage),
(i)-[]-(at:Attacktype),
(i)-[]-(ct:CrimeType),
(i)-[]-(m)
WHERE   p.description =~'(?i).*Al-Qaida.*'
AND wt.description= "Explosives"
AND cntr.name = "Iraq"
AND ct.description =~"(?i).*Insurgency.*"
RETURN i,p,m,w,wt,ws,t,tt,ts,corp,c,pr,r,cntr,n,ctn,d,at,ct
LIMIT 100
```

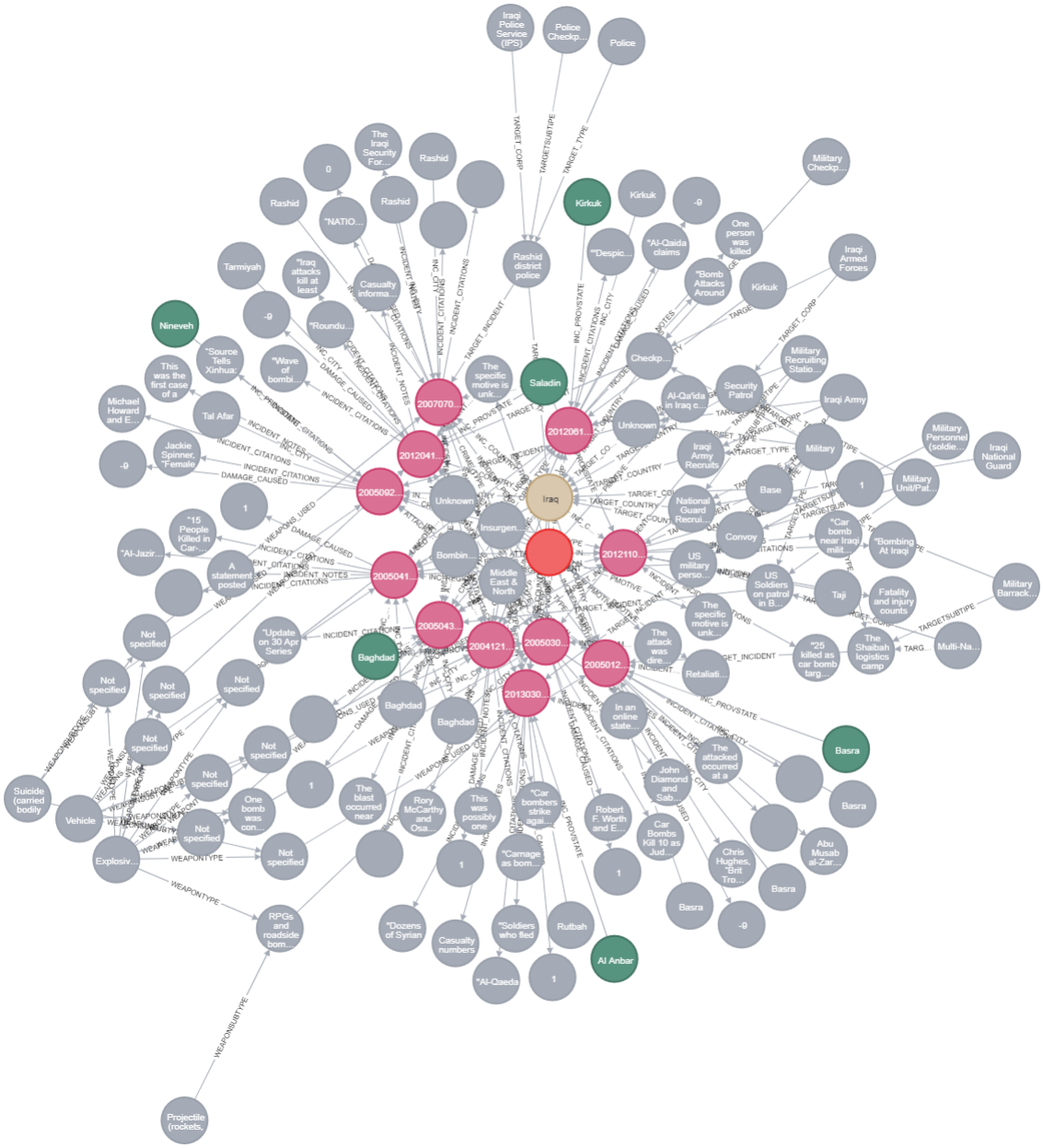


Figure 19: GTD All the information (nodes) conneceted in some way to Al-Qaeda

9. Total number of incidents in Iraq

Query:

```
MATCH (i:Incident)-[r:INC_COUNTRY]-(c:Country)
WHERE c.name="Iraq"
RETURN count (i);
```

	count (i)
1	26057

Figure 20: GTD Total number of incidents in Iraq

5 Conclusions and Future Work

This project, in its entirety, has showcased the added value traditional security domains get, when merging with new technologies. Knowledge graphs in particular can become a valuable tool for researchers, national and transnational law enforcement agencies, as well as, policy makers. The ability to graphically depict known relationships can help for the uncovering of patterns, otherwise unattainable.

The use of Knowledge Graphs in the area of Homeland Security can yield very interesting results for law enforcement and policy makers. This thesis contributes a basic model that can be further enhanced. First of all, the next step is to insert in the ontology and the graph more information, accessible to law enforcement agencies, thus leading the graph to become richer in information. Besides that, Machine Learning Models can be leveraged in order to find patterns in the graph, that can lead to a more effective way of countering terrorist threats, both at the political and operational level.

Bibliography

Awan, J., & Memon, S. (2016). Threats of cyber security and challenges for Pakistan. In International Conference on Cyber Warfare and Security. *Academic Conferences International Limited*.

Brass, I., & Sowell, J. H. (2020). Adaptive governance for the Internet of Things: Coping with emerging security risks. *Regulation and Governance, June*, 11–20. https://doi.org/10.1111/rego.12343

Collarana, D., Galkin, M., Lange, C., Scerri, S., Auer, S., & Vidal, M. E. (2018, October). Synthesizing Knowledge Graphs from Web Sources with the MINTE+Framework. *International Semantic Web Conference* , pp. 359-375.

Gardner, T. (2014). The Promise and Peril of the Anti-Commandeering Rule in the Homeland Security Era: Immigrant Sanctuary as an Illustrative Case. *Louis U. Pub. L. Rev., 34*.

Gruenewald, J., Allison-Gruenewald, K., & Klein, B. (2015). Assessing the attractiveness and vulnerability of eco-terrorism targets: A situational crime prevention approach. *Studies in Conflict & Terrorism, 38(6)*, pp. 433-455.

Haynes, M., & Giblin, M. (2014). Homeland security risk and preparedness in police agencies: The insignificance of actual risk factors. *Police Quarterly, 17(1)*, pp. 30-53.

Hiemstra, N. (2014). Performing homeland security within the US immigrant detention system. *Environment and Planning D: Society and Space, 32(4)*, pp. 571-588.

Hu, H., Zhang, H., Liu, Y., & Wang, Y. (2017). Quantitative method for network security situation based on attack prediction. *Security and Communication Networks*.

Iannacone, M., Bohn, S., Nakamura, G., Gerth, J., Huffer, K., Bridges, R., . . . Goodall, J. (2015, April). Developing an ontology for cyber security knowledge graphs. *Proceedings of the 10th Annual Cyber and Information Security Research Conference*.

Jones, C., Bridges, R., Huffer, K., & Goodall, J. (2015, April). Towards a relation extraction framework for cyber-security concepts. *Proceedings of the 10th Annual Cyber and Information Security Research Conference*, pp. 1-4.

Jung, K., & Park, H. (2014). Citizens' social media use and homeland security information policy: Some evidences from Twitter users during the 2013 North Korea nuclear test. *Government Information Quarterly, 31(4)*, pp. 563-573.

Kahan, J. (2015). Resilience redux: Buzzword or basis for homeland security. *Homeland Security Affairs, 11*, pp. 1-19. Retrieved from https://www.hsaj.org/articles/1308

Kaunert, C., & Leonard, S. (2019). The collective securitisation of terrorism in the European Union. *West European Politics, 42(2)*, pp. 261-277.

Kaynar, K. (2016). A taxonomy for attack graph generation and usage in network security. *Journal of Information Security and Applications, 29*, pp. 27-56.

Kelarestaghi, K., Heaslip, K., Khalilikhah, M., Fuentes, A., & Fessman, V. (2018). Intelligent transportation system security: hacked message signs. *SAE International Journal of Transportation Cybersecurity and Privacy, 1(11-01-02-0004)*, pp. 75-90.

Koulas, E. (2019). *Defining Sovereignty and National Interest on Cyberspace : National and Supranational Paradigms*. Univeristy of Macedonia.

Koulas, E., Anthopoulos, M., Grammenou, S., Kaimakamis, C., Kousaris, K., Panavou, F., Piskioulis, O., Shah, S. I. H., & Peristeras, V. (2020). Misinformation and its stakeholders in Europe : a web-based analysis. *ArXiv Preprint*, 1–35.

Lafree, G., Dugan, L., Fogg, H. V, & Scott, J. (2006). *Building a global terrorism database*. *NCJ 214260*, 208.

http://www.ncjrs.gov/App/Publications/abstract.aspx?ID=235792

LaFree, G., Dugan, L., & Miller, E. (2014). Putting terrorism in context: Lessons from the global terrorism database. In *Putting Terrorism in Context: Lessons from the Global Terrorism Database*. https://doi.org/10.4324/9781315881720

Le, N., & Hoang, D. (2016, December). Can maturity models support cyber security? *IEEE 35th international performance computing and communications conference (IPCCC)* , pp. 1-7.

Miehling, E., Rasouli, M., & Teneketzis, D. (2015, October). Optimal defense policies for partially observable spreading processes on Bayesian attack graphs. *Proceedings of the Second ACM Workshop on Moving Target Defense*, pp. 67-76.

Noel, S., Harley, E., Tam, K., Limiero, M., & Share, M. (2016). CyGraph: graph-based analytics and visualization for cybersecurity. In *Handbook of Statistics (Vol. 35)* (pp. 117-167). Elsevier.

Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., & Taylor, J. (2019). Industry-scale knowledge graphs: lessons and challenges. *Queue, 17(2)*, pp. 48-75.

Papadaki, S., Banias, G., Achillas, C., Aidonis, D., Folinas, D., Bochtis, D., & Papangelou, S. (2017, July). A humanitarian logistics case study for the intermediary phase accommodation center for refugees and other humanitarian disaster victims. In *International Conference on Dynamics of Disasters* (pp. 157-202). Springer, Cham.

RAND Corporation. (n.d.). *RAND Database of Worldwide Terrorism Incidents*. Retrieved October 4, 2020, from https://www.rand.org/nsrd/projects/terrorism-incidents.html

Sageman, M. (2014). The stagnation in terrorism research. *Terrorism and political violence, 26(4)*, pp. 565-580.

University of Maryland. (2019). GTD Codebook: Inclusion Criteria and Variables. *National Consortium for the Study of Terrorism and Responses to Terrorism (START), October*, 65. http://www.start.umd.edu/gtd/downloads/Codebook.pdf

Xia, T., & Gu, Y. (2019, July). Building terrorist knowledge graph from global terrorism database and wikipedia. *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 194-196.

Appendix I

The whole script used for constructing the RANDWTI graph. It is important to be run all at once.

```
LOAD CSV WITH HEADERS FROM "file:///RAND1.csv" AS row
MERGE (pName: Perp_Name {name: row.Perpetrator})
MERGE (Country: Target_Country {name: row.Country})
MERGE (City: Target_City {name: row.City})
MERGE (Incident: Incident {name: row.ID})
MERGE (pWeapon: Perp_weapon {name: row.Weapon})
MERGE (Description: Incident_Desc {name: row.Description})
MERGE (Damage: Incident_Damage {Fatalities: row.Fatalities, Injuries: row.Injuries})

MERGE (pName)-[:INCIDENT_PERP]->(Incident)
MERGE (pName)-[:ATTACKED]->(Country)
MERGE (pWeapon)-[:WEAPON_USED]->(Incident)
MERGE (Country)-[:INCIDENT_COUNTRY]->(Incident)
MERGE (City) - [:INCIDENT_CITY]->(Incident)
MERGE (Incident)-[:DAMAGE_CAUSED]->(Damage)
MERGE (Incident)-[:INCIDENT_DESC]->(Description)
MERGE (City)-[:IS_IN]->(Country)
```

Appendix II

The whole script used for constructing the GTD graph.

```
LOAD CSV WITH HEADERS FROM "file:///gtd_incidents.csv" AS line
WITH line WHERE line.eventid is not null
CREATE (i:Incident {eventid: toFloat(line.eventid), year: toInteger(line.iyear), month:toInteger( line.imonth), day:toInteger( line.iday), approxdate: line.approxdate, summary: line.summary, latitude: toFloat(line.latitude),longitude: toFloat(line.longitude), country: line.country_txt,region: line.region_txt,provstate: line.provstate,city: line.city});

CREATE INDEX ind_eventid FOR (i:Incident) ON (i.eventid);
CREATE INDEX ind_inc_country FOR (i:Incident) ON (i.country);
CREATE INDEX ind_inc_region FOR (i:Incident) ON (i.region);
CREATE INDEX ind_inc_provstate FOR (i:Incident) ON (i.provstate);
CREATE INDEX ind_inc_city FOR (i:Incident) ON (i.city);

LOAD CSV WITH HEADERS FROM "file:///gtd_incidents.csv" AS line
WITH line WHERE line.eventid is not null
CREATE (i:Incident {eventid: toFloat(line.eventid), year: toInteger(line.iyear), month:toInteger( line.imonth), day:toInteger( line.iday), approxdate: line.approxdate, summary: line.summary, latitude: toFloat(line.latitude),longitude: toFloat(line.longitude), country: line.country_txt,region: line.region_txt,provstate: line.provstate,city: line.city});

CREATE INDEX ind_eventid FOR (i:Incident) ON (i.eventid);
CREATE INDEX ind_inc_country FOR (i:Incident) ON (i.country);
```

```
CREATE INDEX ind_inc_region FOR (i:Incident) ON (i.region);
CREATE INDEX ind_inc_provstate FOR (i:Incident) ON (i.provstate);
CREATE INDEX ind_inc_city FOR (i:Incident) ON (i.city);

LOAD CSV FROM "file:///city.csv" AS row
CREATE (c:City {name: row[0], provstate: row[2], country: row[1], region: row[3]} );

CREATE INDEX ind_city_name FOR (c:City) ON (c.name);

MATCH (i:Incident)
MATCH (c:City {name: i.city} )
CREATE (c)-[r:INC_CITY]->(i);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH row where row.provstate is not null
WITH toFloat(row.eventid) AS reventid, row.provstate AS rprovstate,  row.country_txt AS rcountry, row.region_txt AS rregion
MATCH (i:Incident{ eventid: reventid})
MERGE (p:Provstate {name: rprovstate ,country:COALESCE(rcountry, "") ,region:COALESCE(rregion, "") })
MERGE (p)-[r:INC_PROVSTATE]->(i);

CREATE INDEX ind_provstate_name FOR (p:Provstate) ON (p.name);

LOAD CSV FROM "file:///country.csv" AS row
CREATE (c:Country {name: row[0], region: row[1]} );

CREATE INDEX ind_country_name FOR (c:Country) ON (c.name);

MATCH (i:Incident)
```

```
WHERE i.country is not null
MATCH (c:Country {name: i.country} )
CREATE (c)-[r:INC_COUNTRY]->(i);

LOAD CSV FROM "file:///region.csv" AS row
CREATE (r:Region {name: row[0]} );


CREATE INDEX ind_region_name FOR (r:Region)ON (r.name);

MATCH (i:Incident)
WHERE i.region is not null
MATCH (r:Region {name: i.region} )
CREATE (r)-[rel:INC_REGION]->(i);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH  row.attacktype1_txt AS attacktype1
MERGE (a:Attacktype {description: attacktype1} );

CREATE INDEX ind_attacktype FOR (r:Attacktype)ON (r.description);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH row where row.attacktype1_txt is not null
WITH  toFloat(row.eventid) AS reventid, row.attacktype1_txt AS attacktype1_txt
MATCH (i:Incident {eventid: reventid })
MATCH (a:Attacktype {description: attacktype1_txt })
CREATE (a)-[r:ATTACK_USED]->(i);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
```

```
WITH row where row.alternative_txt is not null
WITH  toFloat(row.eventid) AS reventid, row.alternative_txt AS alternative_txt
MERGE (c:CrimeType {description: alternative_txt });

CREATE INDEX ind_crime FOR (r:CrimeType)ON (r.description);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH row where row.attacktype1_txt is not null
WITH  toFloat(row.eventid) AS reventid, row.alternative_txt AS alternative_txt
MATCH (i:Incident {eventid: reventid })
MATCH (c:CrimeType {description: alternative_txt })
CREATE (c)-[r:CRIME_TYPE]->(i);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH  toFloat(row.eventid) AS reventid,row.nkill AS nkill, row.nwoundus AS nwoundus, row.property AS property, row.propextent_txt AS propextent_txt, row.propvalue AS propvalue, row.ishostkid AS ishostkid,row.nhostkid AS nhostkid
CREATE (d:Damage {nkill: nkill,nwoundus:nwoundus,property:property,propextent_txt:propextent_txt,propvalue:propvalue ,ishostkid:ishostkid, nhostkid:nhostkid, eventid: reventid });

CREATE INDEX ind_damage_event_id FOR (c:Damage) ON (c.eventid);

MATCH (i:Incident)
MATCH (d:Damage {eventid:i.eventid})
CREATE (i)-[rel:DAMAGE_CAUSED]->(d);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH  toFloat(row.eventid) AS reventid,  row.target1 AS rtarget, row.natlty1_txt AS targetCountry, row.targtype1 as rtargtype1
```

```
MATCH (i:Incident {eventid: reventid })
CREATE (t:Target { targetid:rtargtype1, description:COALESCE(rtarget, ""), country:COALESCE(targetCountry, "") } )
CREATE (t)-[rel:TARGET_INCIDENT]->(i);

MATCH (t:Target),(c:Country)
WHERE t.country = c.name
MERGE (t)-[r:TARGET_COUNTRY]->(c);

CREATE INDEX ind_target FOR (r:Target)ON (r.description);
CREATE INDEX ind_targetid FOR (r:Target)ON (r.targetid);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH  row.target1 AS rtarget , row.targtype1_txt AS rtargettype, row.targtype1 As rtargettypeid
MERGE (tt:TargetType {description: rtargettype, targettypeid:rtargettypeid  });


MATCh (t:Target)
MATCH (tt:TargetType )
WHERE t.targetid = tt.targettypeid
MERGE (tt)-[rel:TARGET_TYPE]->(t);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH row where row.targsubtype1_txt is not null
WITH  row.targsubtype1_txt AS rtargetsubtype, toFloat(row.eventid) AS reventid
MATCH (i:Incident {eventid: reventid })<-[r:TARGET_INCIDENT]-(t:Target )
MERGE (tt:TargetSubtype {description: rtargetsubtype })
MERGE (tt)-[rel:TARGETSUBTIPE]->(t);
```

```
LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH row where row.corp1 is not null
WITH  toFloat(row.eventid) AS reventid, row.corp1 AS rcorp1
MATCH (i:Incident {eventid: reventid })<-[r:TARGET_INCIDENT]-(t:Target )
MERGE (c:Corp1 {description: rcorp1 })
MERGE (c)-[rel:TARGET_CORP]->(t);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH row where row.addnotes is not null
WITH  toFloat(row.eventid) AS reventid, row.addnotes AS raddnotes
MATCH (i:Incident {eventid: reventid })
MERGE (n:Notes {description: raddnotes })
MERGE (i)-[r:INCIDENT_NOTES]->(n);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH row where row.scite1 is not null
WITH  toFloat(row.eventid) AS reventid, row.scite1 AS scite1
MATCH (i:Incident {eventid: reventid })
CREATE (c:Citation {description: scite1 })
CREATE (i)-[r:INCIDENT_CITATIONS]->(c);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH row where row.scite1 is not null
WITH  toFloat(row.eventid) AS reventid, row.scite2 AS scite2
MATCH (i:Incident {eventid: reventid })
CREATE (c:Citation {description: scite2 })
CREATE (i)-[r:INCIDENT_CITATIONS]->(c);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
```

```
WITH row where row.scite1 is not null
WITH  toFloat(row.eventid) AS reventid, row.scite3 AS scite3
MATCH (i:Incident {eventid: reventid })
CREATE (c:Citation {description: scite3 })
CREATE (i)-[r:INCIDENT_CITATIONS]->(c);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH  toFloat(row.eventid) AS reventid, row.weaptype1_txt AS weaptype, row.weapdetail As  weapdetail
MATCH (i:Incident {eventid: reventid })
CREATE (w:Weapon {description: weaptype , description: COALESCE(weapdetail, "Not specified"), eventid :reventid  })
CREATE (w)-[r:WEAPONS_USED]->(i);

CREATE INDEX ind_weapon_inciden_id FOR (w:Weapon) ON (w.eventid);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH row where row.weaptype1_txt is not null
WITH  toFloat(row.eventid) AS reventid, row.weaptype1_txt AS weaptype
MATCH (w:Weapon {eventid: reventid })
MERGE (wt:WeaponType {description: weaptype  })
MERGE (wt)-[r:WEAPONTYPE]->(w);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH row where row.weapsubtype1_txt is not null
WITH  toFloat(row.eventid) AS reventid, row.weapsubtype1_txt AS weapsubtype
MATCH (w:Weapon {eventid: reventid })
MERGE (ws:WeaponSubtype {description: weapsubtype  })
MERGE (ws)-[r:WEAPONSUBTYPE]->(w);
```

```
LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH  toFloat(row.eventid) AS reventid, row.gname AS description
MATCH (i:Incident {eventid: reventid })
MERGE (p:Perpetrator {description:description  })
MERGE (p)-[r:INCIDENT_PERP]->(i);

CREATE INDEX perp_name FOR (p:Perpetrator) on (p.description );

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH row where row.motive is not null
WITH  toFloat(row.eventid) AS reventid, row.motive AS rdescription
MATCH (i:Incident {eventid: reventid })
MERGE (m:Motive {description: rdescription  })
MERGE (m)-[r:INCIDENT_MOTIVE]->(i);

LOAD CSV  WITH HEADERS FROM "file:///gtd_incidents.csv" AS row
WITH row where row.motive is not null
WITH  toFloat(row.eventid) AS reventid, row.motive AS motive, row.gname AS pdesc
MATCH (i:Incident {eventid: reventid })
MATCH (m:Motive {description: motive })
MATCH (p:Perpetrator  {description: pdesc })
MERGE (p)-[rel:PMOTIVE]->(m);
```