USING AUTOMATED KEYWORD EXTRACTION TO FACILITATE TEAM DISCOVERY IN A DIGITAL FORENSIC INVESTIGATION OF ELECTRONIC COMMUNICATIONS

W.J.C. van Staden* **and E. van der Poel**†

* *School of Computing, UNISA Science Campus, Florida Park, South Africa. E-mail: wvs@wvs.za.net*
† *School of Computing, UNISA Science Campus, Florida Park, South Africa. E-Mail: evdpoel@unisa.ac.za*

Abstract: A major problem that often occurs in Digital Forensics (DF) is the huge volumes of data that has to be searched, filtered, and indexed to discover patterns that could lead to forensic evidence. The nature of, and the process by which the data gets collected, implies that the data also contain information about persons that are not implicated, or only incidentally involved in the crime under investigation. Privacy is therefore an important issue that needs to be managed in a DF investigation. This paper shows that techniques used in the Team Formation (TF) task can be successfully applied to address both the problems of data volume and privacy. The TF task can be re-formulated to fit the DF arena: to commit a crime, the culprit(s) may require the assistance of several other individuals, which implies that a team of some sort gets established. During a post-mortem DF analysis, an investigator may only have one, or a few names to start with. One of the key challenges is finding possible co-conspirators. From a TF point of view, the culprit is trying to find the best team to commit the crime, given some constraints. The TF task in DF requires the recording of skill-sets, and the generation and/or discovery of a graph depicting interaction between candidates. If the data consist of an email corpus and peoples' roles in an organisation (such as in the Enron data), both of these are readily available. In this paper we consider the TF problem in general and extend it to the DF arena by considering the information that an investigator may have access to during the investigation. We also show that simple information retrieval and keyword extraction techniques (such as RAKE) can be used to automatically discover potential teams from the data, while preserving privacy; results from a series of experiments (using the new definitions of TF and the proposed information retrieval techniques) on the Enron data is then presented.

Key words: Digital Forensics, Digital Forensic Investigation, Cyber-crime, Team-formation, Social Network Analysis, Expert Finding

1. INTRODUCTION

The post-mortem forensics analysis of communications data, such as an email corpus can be an extremely difficult and time-consuming task due to the volume and weakly structured nature of the data. The analysis process usually involves a traditional brute-force search for specific patterns, filtering to reduce the search space, and indexing of the resulting documents, or parts of documents. The patterns, filters and indexing mechanisms are often hand-crafted by the investigator, usually specific to the potential crime being investigated. That is, the use of 'hunches' provide the initial stepping stones for an investigative effort during the early stages.

Proposed techniques use machine learning [1] and data-mining techniques [2] to guide the investigator's efforts by highlighting 'low-hanging fruit'. These techniques and tools save time and allow the investigator to more quickly find results that could lead to evidence. Another idea would be to explore the data to find possible teams within the forensic data.

The creation and formation of teams have been studied in operations research and the management and social sciences. In operations research the Team Formation (TF) problem consists of optimally assigning people with certain skills to a task to be accomplished, for example building a software development team. In the social sciences the TF techniques often are used to do a post-hoc discovery of teams, by using individuals' communication patterns. Graphs are constructed from these patterns showing communication habits and patterns – but the focus is not necessarily on the ability of such persons to form a team around a particular set of tasks.

Crimes often involve the creation of teams, where a team would not be as rigid and designed as in the case of a software development team. Such a team is likely to be sub-optimal from a skills perspective, as there would be the additional constraint that the potential team member would have to be willing, or be able to be coerced to commit acts that would assists in the crime. There may even be unwitting team members, who participates in the crime through the simple act of doing their jobs. The TF task in the planning and execution of a crime therefore has possible additional dimensions. This also implies that a team may not necessarily all be aware of the crime being committed – thus the construction of a team could potentially include members who are simply used as 'tools' in order to commit the crime.

This paper shows that techniques used in TF discovery

Based on: "Team Formation in Digital Forensics", by W.J.C. van Staden and E. van der Poel which appeared in the Proceedings of Information Security South African (ISSA) 2016, Johannesburg, 17 & 18 August 2016. © 2016 IEEE

can be applied to the Digital Forensics (DF) task to automatically discover potential teams involved in the crime. This means that the investigator has a much smaller set of potential culprits to start investigating, than using more traditional investigation techniques. It also has the benefit that the investigator does not need to look at the data of potentially innocent persons whose data happens to form part of the corpus. This has positive implications for privacy.

The TF problem is therefore considered from the perspective of the culprit(s); if someone wanted to commit a crime, who would the best team be to accomplish this? The word 'team' should be considered a loose term, as the team may involved people who are simply doing their normal jobs, or may involve people, who has information required to accomplish aspects of the crime, and may or may not know that they are providing the information to aid in the commission of a crime.

Applying TF techniques can be viewed as intelligent automated filters that aim to (hopefully substantially) reduce the list of potential suspects. As in any investigation, these persons should remain 'just' suspects until further corroborating evidence is found.

To illustrate the concepts of applying TF discovery in DF, the Enron email corpus[*] was used as the data under investigation. Since the Enron data-set has undergone several releases in which data has been removed (at the request of persons whose data was within the data-set) the data provided can no longer be used to identify those who were indicted, implicated, or sentenced – hence, for the moment, we cannot provide error rates or accuracy (recall and precision), however, it is important to understand that the purpose of the proposed techniques is not to provide an automated system for solving cyber-crime – the purpose is to provide tools and techniques that can guide an investigator through the investigation, and importantly, potentially protect the privacy of parties that may not be involved in the crime.

1.1 Contribution

This paper contributes to the field of Digital Forensics (DF) by applying techniques of the Team Formation (TF) task from a digital forensic perspective. It is argued that the TF task can be applied during a post-mortem analysis of seized data to guide the investigator, by narrowing down the list of suspects, focusing on persons of immediate interest, and avoiding investigating potentially innocent persons. To facilitate the use of TF, however, the team formation task has to be placed in the correct context.

In general, TF considers social network graphs and potential team members' skills and expertise to build a team to complete a specific task. The important difference between this work and others is that the team formation problem is framed in the DF paradigm, specifically with the focus on guiding the investigator during the analysis. It is shown that standard Information Retrieval (IR) techniques can be employed to extract information from an email corpus, that can lead to identifying teams. The formulation of the TF task in the DF paradigm will allow further research into automation of the guidance provided to the investigator. A formal notation for the TF task is also proposed. This notation can be used when reasoning about the team formation problem in this and future research.

We further show that by using automated keyword extraction (also a technique from the IR field) that TF can be further aided by identifying potentially telling keywords and phrases that identify persons within the corpus.

Additionally, by allowing the investigator to focus specifically on persons of interest (i.e those in the team), the privacy of others whose data forms part of the seized data may be protected.

1.2 Structure of the paper

The rest of the paper is structured as follows:

- Section 2. provides background information on DF, the TF task and related work.
- Section 4. frames the TF task in the DF paradigm, and provides formal definitions for ranking individuals.
- Section 5. provides some examples of the application of the ideas presented in the paper to the Enron mail corpus.
- Section 5.2 discusses the use of automated keyword extraction techniques in order to explore the Enron mail corpus.
- Section 6. presents a discussion on the techniques applied in this paper.
- Finally, section 7. provides concluding remarks.

2. BACKGROUND AND RELATED WORK

Reformulation of the team formation problem concerns itself with two important pieces of work. Firstly, DF provides the paradigm within which the problem is contextualised, secondly, the team formation problem provides the concepts and tools needed to reformulate and understand the problem. Each of these is discussed in turn in the following sections. In order to illustrate the potential use of the work, the use of IR and keyword extraction techniques is also presented.

2.1 Digital Forensics

Digital Forensics (DF) is defined as the "...preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of evidence in a digital context [3]." Using sound forensic techniques and proper controls digital data that could potentially be

[*] The Enron corpus was downloaded from http://tinyurl.com/myjmcjl

evidence is gathered, analysed and presented in context as part of the cyber-crime investigation. Politt [4] calls this the creation of a narrative.

This paper is concerned with digital evidence in the form of data. In particular, the post-mortem analysis (as opposed to live analysis) of de-obfuscated data. Since data can be hidden, a lot of DF research goes into the finding and identification of data. These techniques involve file-carving to find deleted data [5, 6], similarity hashes to identify files or parts of files [7, 8], to name but two.[**] Once data has been de-obfuscated, that is, their meaning can be readily inferred, an analysis on the content can be done which will contribute to the narrative.

The analysis of the data can also be seen as a de-obfuscating effort (since data is now added to the narrative, and therefore its meaning in the narrative becomes clear). However, this paper will stick to the term analysis in order to avoid confusion.

Sifting through large volumes of data is typically accomplished through brute force approaches in which strings of data are matched against search queries, or where meta-data is matched against search queries. Such meta data consists of file-types, time-stamps, file-ownership and so on. Fei et al. [1] propose the use of Self-Organising Maps (SOMs) [9] to guide the investigator. Their technique uses meta-data to detect anomalies in the data, and the investigator is thus guided by focussing analysis on those pieces of data. Fahdi et al [10] also employs SOMs for automated discovery of potential evidence.

Beebe has proposed the use of text-mining to achieve better retrieval rates [11] and as a way to search through large corpora [2], and Pollitt has shown that Natural Language Processing (NLP) techniques such as Named Entity Extraction (NEE) can be useful during the creation of the narrative [4].

The use of automated guidance during a forensic investigation is therefore well established, and this paper builds on those ideas.

2.2 Expert finding and Team Formation

Finding experts is the problem of identifying individuals who may hold knowledge. This particular problem dates back as far as the 1990s [12], and the particular challenge set by the text-retrieval conference (TREC) in 2005 set the scene for renewed research in the field [13].

The particular problem in expert finding is estimating the expertise of an individual. Most notable approaches [12, 14] use a probability distribution model in order to estimate the expertise level. Zhang et al. [15] proposes a propagation based approach to finding an expert within a social network.

The use of social graphs to find criminal associations has been studied by Xu et al [16]. They use shortest-path algorithms to identify associations in criminal networks. However, their evaluation is run purely on the associativity of the links in the network.

Once an expert is found, a social graph is typically used to establish a team of experts within the graph. Team formation is a well researched problem outside digital forensics. Lappas et al [17] make use of minimum-span trees to build a team of experts on topics within a social graph. They show that constructing such a structure is NP-Hard.

Rangapuram et al. [18] extend team formation as presented by Lappas et al to include budget and location constraints. They also allow an upper bound on the team size, and well as a constraint to indicate the minimum level of expertise required to complete the task the team is identified for.

Rahman [19] considers the team formation problem from an economic perspective, and the concept of opaque and translucent teams are introduced. An opaque team shares knowledge within the team in order to maximise the operation of the team. In a translucent team, some information may purposefully remain hidden in order to enhance the attractiveness of the team. Such translucent teams, although not part of this paper, may provide an interesting topic of study once the team formation problem in the DF sphere is well defined.

3. AUTOMATED KEYWORD EXTRACTION

Keyword extraction is the action of scanning text documents with the explicit goal of finding keywords that describe the document under consideration. The simplest technique in this field is the use of single keywords, or using n-grams. Beliga et al. [20] provide a good overview of automated extraction techniques. Different techniques that yield good results such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) [21], can also be used, however, we have not tested these on short unstructured emails as present in the Enron corpus, and this is left for future work. We only present three techniques which relate to the focus of the paper.

Using single words as topic keywords relies on a technique called Term-Frequency-Inverse-Document-Frequency (tf-idf) [22]. This technique uses the relative importance of a word as an indicator of the importance within a corpus of documents. It does so by counting the number of times a term appears within a document under consideration, and multiplying that with the number of documents in the corpus that the keyword appears in. The standard formula is given as:

$$k_{d,k} = log(1 + tf_{d,k} \times log(\frac{|D|}{|D_k| + 1}) \qquad (1)$$

The tf-idf metric works well with single word keywords, however, in many cases a key-phrase is more descriptive

[**] The decryption of data is also, of course, part of the de-obfuscation problem.

of the document under consideration. An n-gram based approach works well in this instance; in order to explain the notion of such an n-gram based approach, first the concepts of bi-grams, tri-grams, and so on are explained. A bi-gram (or 2-gram) is a two-word pair – when using bi-grams, the entire document is represented as a collection of two word pairs. For example, a document D consisting of words $(w_1, w_2, \ldots, w_l)$, the bi-gram representation is $((w_1, w_2), (w_2, w_3), \ldots, (w_{l-2}, w_{l-1}), (w_{l-1}, w_l))$. Using this approach it is possible to construct tri-grams, quad-grams, up to a generalised n-gram representation.

The n-gram approach suffers from scalability problems – determining the correct value of n becomes difficult, since a document will have to be scanned multiple times in order to discover $2, 3, 4, ...n$ combinations of words as key phrases in the document. In most cases a bi-gram or tri-gram approach is sufficient in order to grab to most general keywords. A general advantage of the tri-gram approach is that it does not suffer from 'stop word' exclusion in the key-phrase.

A fast approach (Rapid Automated Keyword Extraction (RAKE)) with acceptable results to solving the keyword extraction problem was presented by Rose et al. [23]. RAKE uses word co-occurrences as a way to determine key-phrase boundaries. A document is split into sentences, and each sentence is divided by stop-words. Typical stop-words are words such as determiners, coordinating conjunctions, and so forth. Whatever remains is considered as key-phrases. Using a co-occurrence matrix, each key-phrase is given a weight. In the typical approach two weights are assigned: $deg(w)$ which indicates the number of times the word appears in the document, and $freq$ which is the number of words that appear with w as a key-phrase.

Using stop-words as key-phrase boundaries does result in certain phrases being missed (such as phrases of the form x $of x$). To avoid this, an implementation of the algorithm can implement stop-word spanning which accepts a key-phrase as a legitimate phrase based on some criteria – the original RAKE implementation uses a key-phrase with a stop word if it occurs at least twice in the document.

The following section formulates the TF task in DF.

4. THE TEAM FORMATION PROBLEM IN DIGITAL FORENSICS

Generally speaking, a (cyber-)criminal contemplating a crime has the same problem as a project manager: find a team that will successfully complete a project. The project requires a specific set of skills and/or knowledge related to the task. A project manager aims to find the best group of experts that the budget will afford. All the team members will have full knowledge of their role in the team. On the other hand, the criminal has a more complex notion of 'afford', in that the criminal should be able to convince or influence potential members to commit parts of the crime. This means that the team may well not consist of the 'best' experts. The are also likely to be team 'members' who are not aware of their role in the crime, or even be aware that a crime is being commitment, through the simple execution of their jobs, or sharing of their knowledge. We define 'aid' as either the execution of a specific task, such as a job function, or the sharing of specific knowledge to assist in the execution of specific tasks.

The team formation problem is therefore formulated for DF investigations, as follows:

Definition I The Team Formation Problem in a Digital Forensics Context
Given a set of individuals Ψ, a set of topics they have knowledge about Θ, a graph depicting their communication habits $G =< V, E >$, (where V is a set of vertices representing the individuals and E is a set representing the edge between the vertices from V) and a topical definition of a committed act, find $\Gamma \subset \Psi$ which depicts a likely team needed to either commit the act, or who will be able to provide aid in order for the act to be committed (the graph provides clues to persons who may potentially collude in order to accomplish a specific task). ■

A formal definition of the notation in formulating the team formation problem in the DF context is provided in definition 4.1.

It is important to understand the notion of a 'likely' team. The suspect may not have looked for the most influential people, or all the experts in order to commit a crime, any person who has the knowledge or can lead to knowledge may be sufficient. In particular the criminal may have had individuals in mind who had knowledge, and whom he would be able to influence.

This leads to a paradox in the existing definitions of team formation: teams may not consist of the best choices, and may more than likely resemble *translucent* teams [19] in which the criminal and co-conspirators hold a residual claim on the team. This paradox is defined as follows.

Definition II The Team Formation Problem Paradox
In order to accomplish the task at hand, the cyber-criminal's choice in team may not consist of the experts, or seats of power in the organisation. Normal team formation analysis techniques rely on building a team from influential people or experts, meaning traditional team formation analysis techniques may be of limited use in this case.

Additionally, the suspect may not be part of the team produced during a traditional TF analysis. ■

This does not mean that traditional team formation analysis techniques are useless. Since traditional team formation coupled with Social Network Analysis (SNA) provides valuable information on the potential team that could be formed, they can act as a good guide during the investigative process.

The team formation problem as defined above therefore requires de-obfuscated data from which the following can be derived: a social graph for the persons under investigation, topics extracted from the data, and a framing of the act in terms of the topics. This last concept is important, since the investigator must have enough knowledge of the domain being investigated in order to frame the act in terms of the topics, which leads to the following definition of the act or crime.

Definition III The Crime as a Task
In the team formation problem for cyber-crime an *act*, is a task that can be defined based on knowledge that is required to complete it. Knowledge can be encoded into language phrases, of which several can be used to define the act. ■

Based on the above requirements, the team formation problem is considered with respect to seized email data. The choice of using email data aids in:

1. Constructing a social graph from the email data can be easily automated.
2. Extracting topics from the data can be approximated by performing noun-phrase-, and named entity extraction. Moreover, general IR techniques allows the easy indexing of large email corpora.
3. The terms used to define the act will correspond to the extracted terms and can thus be used during the guided investigation.

The following section considers the the examination of email data.

4.1 Examining Email Corpora

Given the team formation problem as defined in Definition 4., this section considers the identification of what is termed a *candidate team*. This is a team that consists of all the individuals that could potentially form part of an *ideal team*. An ideal team is a team that may have fit the requirements of the suspect.

The Aardvark social search engine [14] attempted to find individuals that may have been able to answer questions from other individuals. It did so by determining the likelihood that a particular individual would be able to answer a question on a certain topic. Aardvark uses NLP techniques, as well as crafted profiles to build its model of users and their ability to answer question on particular topics.

The paper builds on this idea, by showing that an easy approximation for topics, and the social network of the individuals can be used to build a likely team (Definition 4.) for committing the crime.

To accomplish this the following is to be done prior to the analysis phase:

1. Create an index on topics for the corpus,
2. Create a communications network for the users of the mail system,
3. Define the act using nomenclature from the enterprise context,
4. Generate a sub-graph depicting the individuals involved in communication about the topic,
5. Use the sub-graph as a basis for further analysis and investigation.

The set of topics each team member is knowledgeable on is derived through IR techniques from the seized email corpus S.

For any corpus S, the following is defined for the team formation problem in cyber-security:

Definition IV Team formation problem notation
The following notation is defined for the team formation problem:

1. Θ represents all topics embedded in S,
2. $\theta \in \Theta$ is the set of all topics that forms part of a search on S.
3. Ψ represents all the individuals within the corpus,
4. δ_u represents all the documents directly related to individual $u \in \Psi$. Directly related means that this individual has a copy of this document in their possession.
5. $\psi \subseteq \Psi$ is the set of individuals who are under consideration. It may be that certain individuals are excluded from the investigation from the start, therefore, although S may be about Ψ, only the set ψ is under consideration. As the investigation progresses more individuals may be added to Ψ and removed from ψ (or vice versa).
6. δ_u^t is the set of all documents for user u on topic $t \in \Theta$
7. $util(u)$ is a utility rating for u.
8. $G =< V, E >$ is the social graph depicting the interaction between all $u \in \Psi$, with $V \subseteq \Psi$ and $E = \{(u_k, u_j) | u_k, u_j \in V\}$

■

For every individual in S, it is clear that their share of the mail will be a representation of the set of topics they deal with on a daily basis. Having no other information, it is reasonable to assume that this is a reflection of their knowledge on different topics. Consider for example the employee that spends ninety percent of their time corresponding about new contracts. It is reasonable to assume that they have knowledge on contracts and at least some of the process around them. The utility of this individual to the team is thus a function of the probability distribution given for the user given that topic t is discussed.

$$util(u) = p(u_i|t) \tag{2}$$

The utility function is purposefully provided as a function that could be used as part of an objective function calculation. Since 2 can be changed to represent specific constraints. As it stands, equation 2, assumes a steady state – that is, no new information as it becomes available during the investigation is considered. Consider for example a deposition which reveals beyond doubt that a particular individual had knowledge pertinent to the investigation. Thus, the utility function could be modified to reflect this, and the selection of *candidate* team would change. In section 5.2 the utility function is changed to use a RAKE specific calculation in order to present candidate teams and persons of interest.

Searching for the topic $\alpha \in \Theta$, the result corpus $s \in S$ will contain emails exchanged by individuals within the enterprise. Depending on the nature of the topic, the likelihood of an individual u_i corresponding (either receiving or sending an email) on the particular topic is (using Bayes' theorem): $p(u_i|t) = \frac{p(t|u_i)p(u_i)}{p(t)}$.

Since S is available as the sample space, it is easy to calculate $p(t|u_i)p(u_i) = p(u_i \cap t)$. Which in turn is calculated as in equation 3.

$$p(u_i \cap t) = \frac{|\Delta_u^t|}{|S|} \tag{3}$$

Here δ_u^t is the set of all documents covering topic t from individual u (as defined in 4.1), and $|S|$ is the size of the entire corpus.

Individuals can now be ranked based on the utility they could potentially add to the team (since $\sum_{u_i \in \Psi} p(u_i|t) = 1$).

Based on the utility rank and the search result, it is possible to construct $G' =< V', E' >$ where $G' \subseteq G$, with the constraint that $V' \subseteq V$. G' is thus a sub-graph of G which depicts only the correspondence on topics t. From the investigator's view point, G' presents the *candidate team* for aiding in a crime that requires knowledge on the subjects that will come from the individuals in the graph.

The resulting *candidate team* graph G' can then be used in well known social network techniques such as *centrality*, span-tree's to determine teams, and dense sub-graphs. However, at this point, the investigator can simply use the G' to guide the analysis of particular emails that could be evidence.

Now that the concepts behind the team formation problem have been articulated, the following section provides some initial samples in using the generation of G' on the Enron email corpus.

5. EXPERIMENTAL RESULTS

In 2001, the Enron energy company was embroiled in a scandal relating to unlawful and unethical financial practices. Enron basically used complex financial techniques in order to hide their losses, thereby artificially boosting the company's stock value. During the investigation, the email of several hundred of the key employees in Enron was seized and analysed.

Subsequently, the corpus was purchased and released by Andrew McCallum who prepared the content and released the emails in a folder-based hierarchy, all in *mbox* (RFC4155) format [24]. Petitions by several individuals resulted in their emails being removed from the corpus, and the result is a corpus of one-hundred and fifty individuals spanning around 517,000 emails.

There has been a lot of research done on the corpus, including data mining, social network analysis based on the communication links between individuals, and so on. The ideas presented here are (as far as the authors are aware) the first examination of a team formation problem on the Enron corpus – specifically with the team formation problem framed in the DF context.

The purpose of the experiment for this paper was to consider the team formation problem on a real-world set of data. It is shown that very simple techniques can go a long way in providing guidance to the investigator when sifting through volumes of data.

The experiment was conducted based on the steps outlined in section 4.1:

1. The entire email corpus (that was made available) was indexed, and an inverted index was created. This resulted in around 780,000 unique search terms for the 517424 emails all stored in RFC822 *mbox* format.
2. For the communications network (or social network graph) of the persons involved.
3. Several key phrases representing 'topics' were used to search the corpus (thus describing the act in terms of knowledge needed to commit or to aid in committing),
4. A sub-graph of the individuals who communicated about the topics was created, and merged into a graph that represents a *candidate team* for the act. We avoided only listing instances where persons in the candidate team communicated about the topic of interest to allow the full nature of the interaction between individuals to come to light.

We approached the identification of candidate teams from two angles. First, we used simple information retrieval techniques using keyword searches. This allows the investigator to find a candidate team based on a fast keyword search, but requires that some knowledge or ‘hunches’ of potentially interesting keywords are known a priori.

In the second approach we used an automated keyword extractor which provides important keywords within the corpus. This allows the investigator to start their search by using these considered keywords as a starting point.

The information retrieval techniques are discussed next, followed by the automated keyword techniques. Finally the results and findings are presented in section 4.1.

5.1 Information Retrieval Techniques

Some more comments on the information retrieval techniques used are in order. The term dictionary constructed from the corpus contains terms stemmed using the Porter stemmer, and queries run against the term database are stemmed before the search is done. The social network graph for the employees consists of the interaction between Enron employees based on their in-box and sent mail folders.

Although the graph consists of all persons interacting based on the information from the mentioned sources, the visual graphs presented are restricted in two ways: firstly, only individuals from within Enron are displayed on the visualisation, and secondly, based on the likelihood calculation presented in equation 2, only a limited number of individuals are included in the graph. Both of these reasons are purely for a ease of viewing consideration: a visual graph depicting too many vertexes and their links quickly degrades in readability and thus meaning (in printed format). It was thus decided to limit the number of nodes to something that would be meaningful and would be easily digestible.

Figures 1 (page 7) and 2 (page 7) represent a constrained sub-graph for the topics ‘regulation’ and ‘service provider’ (both provide the *utility value* for each individual in parenthesis).

Figure 1 shows several vertexes that are disconnected – this revealed individuals who were corresponding about ‘regulation’ but likely not with parties in Enron.

Lack of space prevents the presentation of all the sub-graphs, however, the *candidate team* graph which includes the topics presented above is provided in Figure 3. The following ‘topics’ were used for the generation: “Federal Energy Regulatory Commission”, “Regulation”, “Audit”, “Contract”, and “Service Provider”.

Just visual inspection of these graphs already provide good clues as to who the individuals with potential knowledge to help with the act are. Knowledge of the structure of the organisation would enable the investigator to follow potential leads – thus the sub-graph can provide guided investigation.

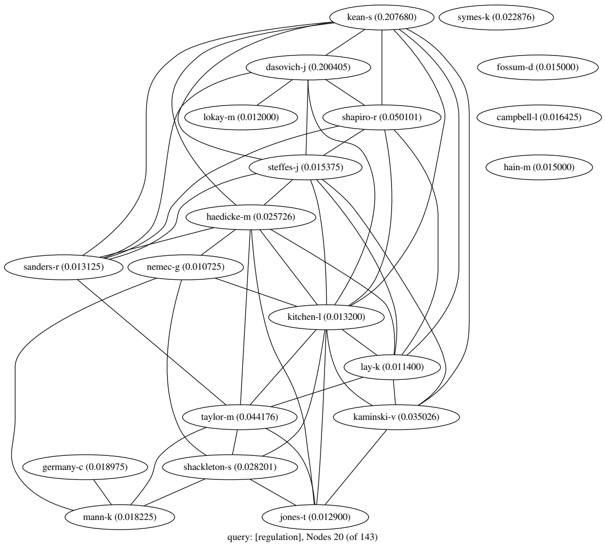

Figure 1: *Candidate Team* for topic ‘regulation’

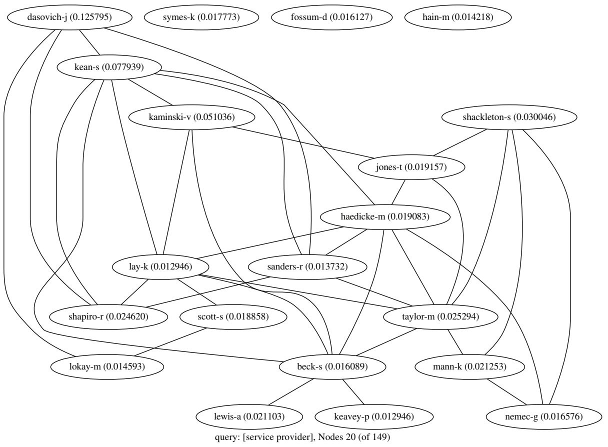

Figure 2: *Candidate Team* for topic ‘service provider’

5.2 Using Automated Keyword Extraction

In this section we consider the use of automated keyword extraction in order to provide potential clues as to good keywords within the corpus. To accomplish this we implemented a basic version of RAKE without stop-word spanning.

We first removed duplicate emails from the corpus, and then scanned each email in order to extract keywords. We also ignored words that contained numbers (as per Lui et al [25] who potentially ignored anything that was not a noun/verb/adjective combination of words). This resulted in approximately 2.4 million keywords and phrases. This also did result in a large number of non-sensical key-phrases (such as a repeating ’a’, and several what can only be described as ‘random keyboard strokes’). Determining if these are noise or material is an interesting topic, and left for future work. Currently, we were only interested in presenting the investigator with keywords that could potentially be of interest – random

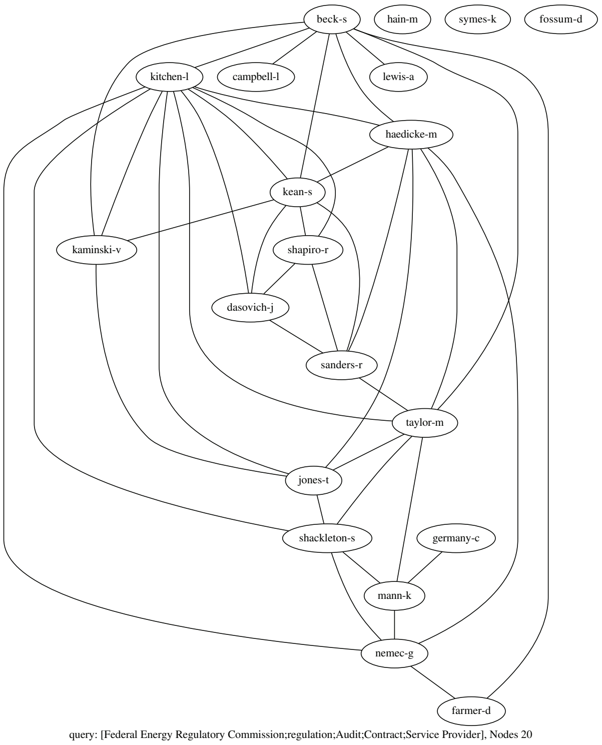

Figure 3: Sub-graph for candidate team for query "Federal Energy Regulatory Commission", "Regulation", "Contract", and "Service Provider"

words like these tended not to show up in the top keyword list for senders.

Keywords were stored in a database which identified the email the phrase appeared in, the employee in whose mail directories the email was found, as well as $deg(w)$ using the consideration from the creators of RAKE (we experimented with a tf-idf index ranking but found no significant advantage using a standard tf-idf approach – further investigation is left for future work).

We followed a similar approach to Rose et al [23] in that only the top third of returned keywords be considered as *extracted*. Once a keyword is not in the top third of keywords for a particular email or user under consideration it is considered *referenced* only. That is, a keyword that is extracted is considered 'extracted' and one that is not is considered 'referenced'.

We used the same techniques as proposed by the creators of RAKE, and calculated the exclusivity, and essentiality of each keyword. All of these use the *extract frequency* ($edf(w)$, and *reference frequency* of a keyword ($rdf(w)$).

The exclusivity of a keyword indicates how often a keyword is extracted when it appears in an email (i.e how often is the keyword in the top third of keywords when it does appear as a keyword):

$$exc(w) = \frac{edf(w)}{rdf(w)} \tag{4}$$

Keywords are then ranked based on 'essentiality'. This is simply an index generated from the exclusivity of a keyword and its reference frequency:

$$ess(kw) = exc(w) \times rdf(w) \tag{5}$$

From the above, we can then easily construct a list of keywords per person, or a global list of keywords that can be used to start digging.

As mentioned earlier, the utility function presented in equation 2 was modified in this approach as equation 6.

$$util(u) = ess(u) \tag{6}$$

6. FINDINGS

The use of automated keyword extraction provided some interesting results which are presented here. The results from the use of information retrieval techniques only provide candidate teams based on the hunches from the investigator, and then only based on simple keywords. Thus the investigator may potentially view data from third parties by following these hunches. Automated keyword extraction tries to reduce the error prone process by extracting important phrases from the corpus, thereby allowing the investigator to focus attention on those phrases and words that make sense from the case point of view. We found some interesting results from our experiments.

Firstly, because keyword extraction uses a statistical model on co-occurrence frequencies, there is no additional information on the semantics of any keywords or phrases that are identified. In an email corpus, this means that the standard phraseology such as *intended recipient*, *confidential information*, and *original message* appears as top-ranked keyword in most profiles of email users within the corpus. This also means that standard platitudes, such as *please find*, *keep well*, *would like* and so on also appear frequently. This is not surprising, since these are standard 'scaffolding' when composing emails which are in essence electronic letter writing.

Secondly, there appears to be strong observational evidence that the principle of Zipf's law [26] applies to the keyword ranking per person from the corpus. Zipf's law states that the frequency of a word in a corpus is inversely proportional to its rank in a frequency table of that corpus. This correlates with the observation above regarding the 'scaffolding' keywords. However, a thorough investigation is left for future work since Zipf's law requires a minimum length document, and emails may not be a proper case for Zipf's law.

Examination of the results of keyword extraction proceeded by choosing a random person from the corpus, and examining the top essential keywords for that person. Generic keywords (as mentioned above) were ignored, and

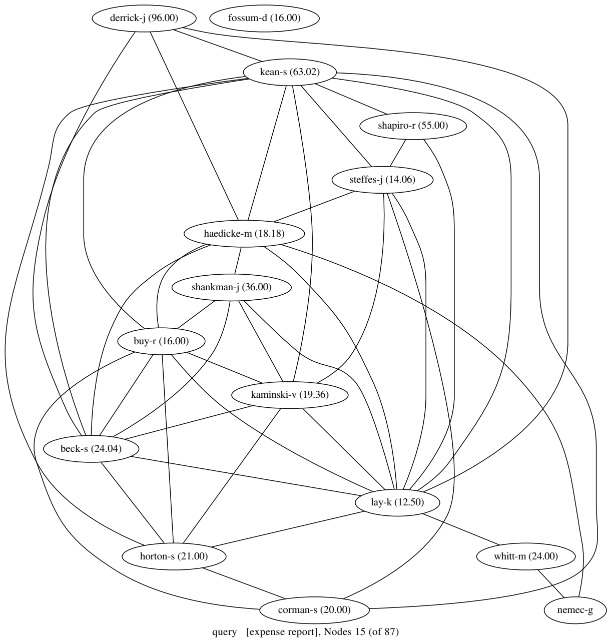

Figure 4: Expense Report

topical keywords that seemed related to the functioning of a large energy corporation were used as anchors for future searches. As an example, the phrase ’expense report’ was discovered using casual inspection (‘expense report’ appears in the top 100 keywords for persons a total of 36 times).

The Figure shows persons in the corpus who communicate, and provides the ‘essentiality’ index for each person for the query under consideration. By examining the top keywords from persons identified in the candidate team in 4, we found “federal energy regulatory commission” as the thirty-first keyword with an essentiality ranking of 83.0 and an exclusivity index of 1.0 (using high index persons in the team). This indicates that in each email this person sent or received this keyword was extracted – thus a highly valuable keyword (see Figure 5 for the candidate team). ‘Federal Energy Regulatory Commission’ appears seven times in the top 100 keywords. Although it occurs far less, its exclusivity index of 1.0 does indicate its importance when it does occur.

Using this approach (viewing top keywords, and inspecting the candidate team) it will be possible to quickly determine which keywords are pertinent to an investigation and which persons are of interest. It is also interesting to note that for Figure 4 there were several candidates who used the keyword ‘expense report’ but did not communicate with any of the other persons in the candidate team. Inspection of the emails revealed that it was a request to approve an expense report related to an employee that was sent to a departmental email address. This also indicates that potentially interesting ‘anomalies’ could be highlighted using the proposed techniques.

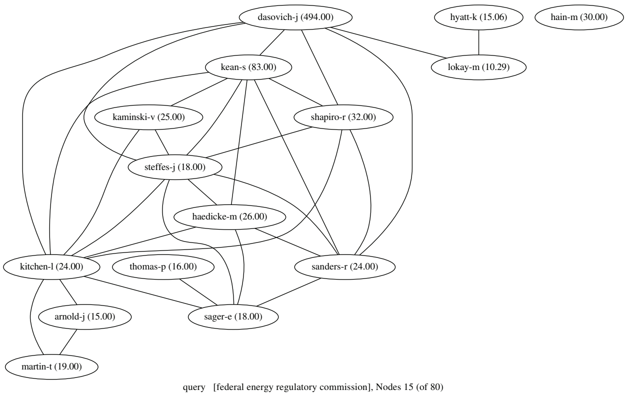

Figure 5: Federal Energy Regulatory Commission (using RAKE)

7. CONCLUSION

This paper reformulates the *team formation problem* within the DF paradigm. Since the team formulation problem is well defined outside of the DF paradigm, it is necessary to place it within the DF context in order to understand it properly. This allows the finer nuances and requirements dictated by the DF paradigm to be understood. In turn, this allows future work to aim specifically at solving particular problems in light of the reformulation. In addition, the team formation problem allows the investigator to be guided by the data within the system. It is important to understand that the proposed techniques should not be considered to be an automated system for solving a cyber-crime, these techniques should only act as a guide for the investigator.

The team formation problem is thus considered from the suspect’s point of view: a crime is defined with respect to topics that are covered by the individuals in the organisation. The team formation problem then identifies the *candidate team* which would likely be able to complete the task (i.e. commit the crime).

This *candidate team* provides the investigator with clues about the individuals within the organisation that may have formed part of the team, or those that may have been used by the suspect in order to complete his task. The important contribution is that the investigator is provided with a guided approach to investigate a large volume of data, thereby focussing the investigation. Additionally, there is an important benefit for privacy of third parties (persons whose emails form part of the seized corpus, but who have nothing to do with the act under investigation). There will be important implications for the investigator and investigation techniques, and further investigation here is also warranted.

The paper also defined formal notations and definitions as the starting point for reasoning and arguing about the team formation problem in the digital forensics perspective. This formal notation can be used as a foundation for future research in this paradigm.

Now that the team formation problem has been formulated for the DF paradigm, it becomes possible to define some future areas of research. These include: using NLP for better topic extraction, such as noun-phrases, or named entities. Once these have been extracted, the investigator can be presented with these ’topics’ as a search filter. Such an approach would mean the investigator no longer needs to carefully craft the search terms, but can rely on the automated system.

We presented results using information retrieval techniques (using simply keyword searches), as well as using automated keyword extraction which extracted phrases and keywords and ranked the relative importance of such phrases as an input to identifying candidate teams.

Future work would also include comparing the results from the techniques proposed herein to regular social network analysis techniques.

Rahman introduced the concept of translucent team [19] in which a team has members that may withhold information from other team members. The effect of such a team within DF would be important to understand, since a cyber-criminal may employ such a team in order to commit a crime – thereby keeping knowledge of the crime away from those who may be able provide evidence.

The prevalence of mentioned ‘scaffolding’ text is a noise removal problem and future work on removing this noise from the keywords (in the automated approach) could significantly reduce the number of interesting keywords.

REFERENCES

[1] B. K. L. Fei, J. H. P. Eloff, M. S. Olivier, and H. S. Venter, “The use of self-organising maps for anomalous behaviour detection in a digital investigation.” *Forensic Sci. Int.*, vol. 162, no. 1-3, pp. 33–7, 2006.

[2] N. Beebe and J. Clark, “Dealing with Terabyte Data Sets in Digital Investigations,” in *Advances in Digital Forensics*. Springer US, 2005, vol. 194, ch. IFIP — The International Federation for Information Processing, pp. 3–16.

[3] G. Palmer, “A Road Map for Digital Forensic Research,” DFRWS, Utica, NY, Tech. Rep., 2001.

[4] M. Pollitt, “History, Histiography, and the Hermeneutics of the Hard Drive,” in *Advances in Digital Forensics IX*, G. Peterson and S. Shenoi, Eds. Seneca, SC, USA: Springer, 2013, pp. 3–19.

[5] N. Alherbawi, Z. Shukhur, and R. Sulaiman, “Systematic Literature Review on Data Carving in Digital Forensics,” *Procedia Technology*, vol. 11, pp. 86–92, 2013.

[6] A. Pal and N. Memon, “The Evolution of File Carving,” *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 59–71, 2009.

[7] J. Kornblum, “Identifying almost identical files using context triggered piecewise hashing,” *Digital Investigation*, vol. 3S, pp. 91–97, 2006.

[8] V. Roussev, “An evaluation of forensic similarity hashes,” *Digital Investigation*, vol. 8, pp. S43–S41, 2011.

[9] T. Kohonen, “The Self Organising Map,” in *IEEE*. IEEE, 1990, pp. 1464–1480.

[10] M. Al Fahdi, N. Clarke, F. Li, and S. Furnell, “A suspect-oriented intelligent and automated computer forensic analysis,” *Digital Investigation*, vol. 18, pp. 65–76, Sep. 2016.

[11] N. L. Beebe and J. G. Clark, “Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results,” *Digital investigation*, vol. 4, pp. 49–54, 2007.

[12] K. Balog, “People Search in the Enterprise,” Ph.D. dissertation, University of Amsterdam, 2008.

[13] A. Bozzon, M. Brambilla, S. Ceri, M. Silvestri, and G. Vesci, “Choosing the Right Crowd: Expert Finding in Social Networks,” in *Proceedings of EDBT/CDT ’13*, ser. EDBT ’13. New York, NY, USA: ACM, 2013, pp. 637–648. [Online]. Available: http://doi.acm.org/10.1145/2452376.2452451

[14] D. Horowitz and S. D. Kamvar, “The Anatomy of a Large-scale Social Search Engine,” in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW ’10. New York, NY, USA: ACM, 2010, pp. 431–440. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772735

[15] J. Zhang, J. Tang, and J. Li, “Expert finding in a social networks,” in *Database Systems for Advanced Applications (DASFAA’2007)*, 2007.

[16] J. J. Xu and H. Chen, “Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks.” *Decision Support Systems*, vol. 38, pp. 473–487, 2004.

[17] T. Lappas, K. Liu, and E. Terzi, “Finding a Team of Experts in Social Networks,” in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’09. New York, NY, USA: ACM, 2009, pp. 467–476. [Online]. Available: http://doi.acm.org/10.1145/1557019.1557074

[18] S. S. Rangapuram, T. Bühler, and M. Hein, “Towards realistic team formation in social networks based on densest subgraphs,” in *WWW 2013*. ACM, 2013, pp. 1077–1088.

[19] D. M. Rahman, “Team Formation and Organization,” Ph.D. dissertation, University of California Los Angeles, 2005.

[20] S. Beliga, “Keyword extraction: a review of methods and approaches,” *University of Rijeka, Department of Informatics, Rijeka*, 2014.

[21] M. Dredze, H. M. Wallach, D. Puller, and F. Pereira, “Generating summary keywords for emails using topics,” in *Proceedings of the 13th international conference on Intelligent user interfaces*. ACM, 2008, pp. 199–206.

[22] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, “Okapi at TREC-3,” 1996, pp. 109–126.

[23] S. Rose, D. Engel, N. Cramer, and W. Cowley, “Automatic keyword extraction from individual documents,” in *Text mining: applications and theory*, M. W. Berry and J. Kogan, Eds. John Wiley & Sons, 2010.

[24] E. A. Hall, “The application/mbox Media Type,” Electronically, September 2005, http://datatracker.ietf.org/doc/rfc4155/.

[25] F. Liu, D. Pennell, F. Liu, and Y. Liu, “Unsupervised approaches for automatic keyword extraction using meeting transcripts,” in *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*. Association for Computational Linguistics, 2009, pp. 620–628.

[26] W. B. Cavnar and J. M. Trenkle, “N-Gram-Based Text Categorisation,” in *SDAIR-94, Third Annual Symposium on Document Analysis and Information Retrieval*, 1994, pp. 161–175.