Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

Using shortest path to discover criminal community

Pritheega Magalingam [a, b, *], Stephen Davis [a], Asha Rao [a]

[a] School of Mathematical and Geospatial Sciences, RMIT University, GPO Box 2476, Melbourne, Victoria 3001, Australia
[b] Advanced Informatics School, Level 5, Menara Razak, Universiti Teknologi Malaysia, Jalan Semarak, 54100 Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 29 January 2015
Received in revised form 9 June 2015
Accepted 22 August 2015
Available online xxx

Keywords:
Criminal network
Shortest path
Leave-one-out
Trust
Suspect
Investigation

ABSTRACT

Extracting communities using existing community detection algorithms yields dense sub-networks that are difficult to analyse. Extracting a smaller sample that embodies the relationships of a list of suspects is an important part of the beginning of an investigation. In this paper, we present the efficacy of our shortest paths network search algorithm (SPNSA) that begins with an 'algorithm feed', a small subset of nodes of particular interest, and builds an investigative sub-network. The algorithm feed may consist of known criminals or suspects, or persons of influence. This sets our approach apart from existing community detection algorithms. We apply the SPNSA on the Enron Dataset of e-mail communications starting with those convicted of money laundering in relation to the collapse of Enron as the algorithm feed. The algorithm produces sparse and small sub-networks that could feasibly identify a list of persons and relationships to be further investigated. In contrast, we show that identifying sub-networks of interest using either existing community detection algorithms or a k-Neighbourhood approach produces sub-networks of much larger size and complexity. When the 18 top managers of Enron were used as the algorithm feed, the resulting sub-network identified 4 convicted criminals that were not managers and so not part of the algorithm feed. We directly validate the SPNSA by removing one of the convicted criminals from the algorithm feed and re-running the algorithm; in 5 out of 9 cases the left out criminal occurred in the resulting sub-network.

Introduction

Retrieving a criminal network from an organised crime incident is an important part of crime investigation. This task is a difficult one, mainly because of the involvement of a variety of criminals who play myriad roles (Basu, 2014; Didimo et al., 2011). In addition to drug trafficking and money laundering, organised crime includes hijacking and equipment smuggling. The task of the criminal investigator is further hampered by the mass of data needing to be searched with an important part of the start of an investigation being the identification of a smaller sample that embodies the relationships within the criminal participants.

In (Magalingam et al., 2014), we presented an algorithm that extracts a small and hence manageable network of e-mail addresses and their relationships from a particular subset of the Enron email dataset; blind carbon copy (BCC) emails with a maximum of two recipients bcc-ed. In this paper we go further. We apply the SPNSA algorithm to larger subsets of the Enron data and we validate the algorithm. We also explicitly compare the size of the subgraphs with those obtained using community detection algorithms and the k-neighbourhood detection methods. The validation of the algorithm was done by dividing the Enron dataset into two

* Corresponding author. School of Mathematical and Geospatial Sciences, RMIT University, GPO Box 2476, Melbourne, Victoria 3001, Australia. Tel.: +61 3 9925 1843.

E-mail addresses: pritheega.magalingam@rmit.edu.au (P. Magalingam), asha@rmit.edu.au (A. Rao).

http://dx.doi.org/10.1016/j.diin.2015.08.002
1742-2876/© 2015 Elsevier Ltd. All rights reserved.

different subsets; 'BCC' and 'TO/CC' email transactions. The first test was to use an algorithm feed not related to the list of known criminals. Note that while this list did contain some criminals, that information was not used. The second test was the 'leave-one-out' test to check whether on dropping a criminal from the algorithm feed the criminal reappears in the resultant network, with success implying that new suspects generated by the algorithm will include any remaining money laundering criminals. The major finding of this paper is that SPNSA is able to solve the difficulties of analysing large and complex networks by using an algorithm feed to form sub-networks without disturbing the structure of the network in the way that community detection algorithms (Pons and Latapy, 2006; Clauset et al., 2004; Newman, 2006; Tasgin et al., 2007) do.

In the past, extracting criminal associations from raw data has required preliminary information of such relationships, and building a network from such, known, relationships has been done manually (Basu, 2014; Didimo et al., 2011; Christin et al., 2010; Oatley and Crick, 2014). For example, Nadji et al. (2013) produce a network of known fraudulent infrastructure by creating links between IP addresses using known attack signatures garnered from passive domain name server and several other sources for malicious activities. Krebs (2002) builds edges between known hijackers of the 9–11 terrorist attacks by manually gathering data from online news articles. The edges, or links, are created based of information such as whether the two persons went to the same school, grew up in the same locality, etc. Oatley and Crick (2014) follow a similar track, using associations such as partner, sibling, cohabitant, to build a relationship network among the members of different UK crime gangs. Clearly, the above methods are time-consuming, and a faster, more automated process of building a relationship network would be very useful for investigators of criminal activities.

We present such an algorithm, which can be run on a large dataset of interactions, to build a more practicable sub-network of known criminals suitable for further investigation. We use the publicly available Enron Dataset (Cohen, 2009), which contains all email communications before and after the collapse of this large company in 2001. This dataset is appropriate for this exercise, as ten people connected with Enron were subsequently convicted of money laundering (Thomsen and Clark, 2004). The structure of the rest of the paper is as follows: In the next section, we describe the Enron dataset in more detail, give the process by which we start the isolation of specific email groupings, compare the connections between the ten criminals in two different email sub-networks, and describe our algorithm. Section 3 gives the results of applying existing community detection algorithms as well as the k-nearest neighbour method, to the Enron dataset to identify the community that the criminals belong. In the section after this, we apply our shortest paths network search algorithm to the two email sub-networks previously identified and compare the results to those obtained by applying the existing community detection algorithms. The penultimate section details the application of our algorithm to the different scenarios that an investigator may encounter. Finally we give the conclusion.

Background

This section describes the preliminary analysis of the Enron email dataset, the people who were convicted of money laundering crime, the identification of criminal communication links and the criminal sub-network formation methods.

Preliminary analysis of dataset

The Enron email dataset contains 1,887,305 email transactions (Cohen, 2009) that were sent using the fields 'TO', 'CC' or 'BCC'. Out of these emails, 16,116 are senders of the emails and 68,203 are receivers of the emails. The Enron email dataset contains a mix of internal and external email transactions. Within the 16,116 email senders, 5831 email transactions are from email addresses that are Enron company email accounts having the name '*enron*' in their email address and the rest of the addresses are external, for example *andrew.fastow@ljminvestments.com*, *anitatr@earthlink.net*, etc. In order to process this large number of emails, we start by extracting the emails sent and received in the last 8 years of Enron – from 1995 to 2002 (Salter, 2008). We clean the data by removing the irrelevant email transactions such as email addresses that have numbers and characters for example '5673@aol.com', that end with airline company name for example '@aircanada.com', that end with 'xpedia.com', 'amazon.com' and other auto response emails.

Several prior works propose ways of extracting criminal networks in the form of associations between texts or people (Basu, 2014; Krebs, 2002). Mining relevant terms from a large volume of police incident summaries and assigning the co-occurrence frequency as a weight to each term is used by (Chen et al., 2004) to design a criminal network while Yang and Ng (2007) use web crawlers to gather identities associated with certain crime related topics in web blog pages and represent them as a network. Similarly, in order to identify criminal cliques, Iqbal et al. (2012) perform chat topic analysis and certain entities that belong to the same chat session are formed into a clique. Louis and Engelbrecht (2011) conduct text mining on passages of a mystery novel to show the association between words, in the form a graph, leading to the identification of murders. In (Anwar and Abulaish, 2012), posts that promote hate and violence in certain dark web forums are grouped in different cliques using an algorithm that measures similarity based on content, time, author and title.

Using keywords as a tool for isolating criminal networks is a problem especially when electronic documents, chat messages, web blogs or emails contain incomplete information or could mislead detection algorithms (Murynets and Piqueras Jover, 2012; Keila and Skillicorn, 2005). Consequently, we choose to ignore the content of the various emails being exchanged between the criminals and propose a very different way to start the building of a criminal network, by considering the type of emails based on recipient fields. As detailed in (Magalingam et al., 2014), we separate the emails with at least one BCC recipient because the existence of a bcc in an email, could indicate a trust relationship (Fox and Schaefer, 2012). While 'to' and