Graph Theory – The Case for Investigating Corruption and Modern Slavery through Suspicious Employment Data

Felicity Gerry
Deakin Univ.
Melbourn, Australia
f.gerry@deakin.com.au

Joseph R. Barr
Acronis SCS
Scottsdale, Arizona, USA
joe.barr@acronisSCS.com

Peter Shaw
Nanjing Univ. of Inf. Science & Technology
JangSu, China
peter.shaw.cs@gmail.com

***Abstract*—This poster uses the mathematics of networks in the novel context of corporate reporting of slavery in supply chains as a method to meet corporate obligations to respect human rights. For those corporates considering risks such as liability for slavery in supply chains, using graph theory, which is capable of sampling affinity in databases, can 'value add' due diligence by scoring identity and veracity.**

***Index Terms*—AI Law, AI Business, Network Models, Modern Slavery, Corruption, Suspicious transactions, Combinatorics, Random Sampling, Graph Theory, Modern Slavery, Human Rights.**

I. Introduction

This poster considers whether graph theory can assist in tackling corruption and modern slavery through identity fraud as a means of corporate reporting of suspicious employment data to compliment suspicious financial reporting. For those corporates considering such risks, using graph theory, which is capable of sampling affinity in databases, can 'value add' due diligence. U.N. Guiding Principles on Business and Human Rights [1] include unequivocal recognition that States have "a duty under international human rights law to protect everyone within their territory and/or jurisdiction from human rights abuses committed by business enterprises". A relatively simple data-mining exercise reduces economic costs and directs those with duties to investigate and report to suspicious employment activity which in turn may identify corruption and slavery in supply chains giving data the potential to be socially meaningful. While extensive research has been done to detect record duplication [2] using statistical methods. Here, graph modelling contributes to existing social value processes. The advantage of using graph theory is that network-based models allow complex many-to-many relationships in the data to be examined. Whereas other Machine Learning and Statistical approaches cannot easily support this. Sampling transactions worthy of further investigation averts the need to analyse a whole database and reduces the impact on data and privacy rights that a full audit might incur. The problem of identifying duplicates is difficult, and a comprehensive analysis requires the entire portfolio, which may well be held across jurisdictions. Analysis on such a scale would be prohibitively expensive. Hence the value of our approach. We consider sampling techniques as a plausible methodology for identifying suspect records.

II. Modelling personas as a graph

A standard procedure involves assigning a node to each vector of identity elements and connecting two nodes if the identity elements agree with some tolerance. A simple modeling will assign weight to every pair of nodes with weight equals to the number of identity elements on which they agree. We agree that no edge is assigned in case the two nodes share no identity element. In the USA for example, a PII consists of five identity elements Social Security, Name (first and last), Address, Phone number and Date of birth, SNAPD in short, five identity elements. So, every vector of length five appearing in the database represents a persona, or a vertex of the persona graph. To disambiguate, we join two vertices and assign a weight, a number between 0 and 5, which equals the number of identity elements on which they agree.

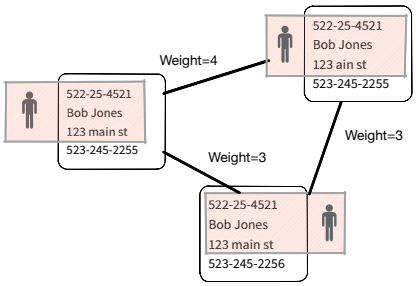

Fig. 1: A graphical depiction of SNAPDs in a database – Is this the same Bob Jones?$(SSN, Name, Address, Phone)$.

Fig. 1 suggests that there are three records which slightly differ indicating the need for investigation of the mainpulations - which are suspicious but may be legitimate, thus there is a streamlined due diligence process. Although it's clearly imperfect, some data scrubbing is possible and is likely to alleviate part of the disambiguation problem. Statistical sampling techniques provides a way to get to the truth without infringing on privacy.

The result is a weighted graph $G = (V, E, w)$ with $w : E \to \{0, 1, 2, \ldots\}$ with a disambiguated database, where all records deemed to represent a single person are identified with an

978-1-7281-8699-3/20/$31.00 ©2020 IEEE
DOI 10.1109/TransAI49837.2020.00025

Authorized licensed use limited to: UNIVERSIDADE DE BRASILIA. Downloaded on November 05,2024 at 19:51:50 UTC from IEEE Xplore. Restrictions apply.

appropriate tag (equivalent to: a database *index*, or *hash*.) . The resulting graph G=(V,E) can be written as the union $\cup T_j$ where T_0 are the isolated vertices, T_1 are the edges, T_2 triangles, and in general T_k is the vertices lying in k-cliques. Clearly if G has l_k k-cliques, then $|T_k| = kl_k$. To illustrate, consider the following (completely plausible) line of reasoning.

Consider a weighted graph (G, w) on n vertices, with an unknown total edge weight W_G. And consider a sample, a sub set D with $k \ll n$ vertices whose weight W_D, i.e., the sum of the weights of all the edges whose endpoint lie in D. Intuitively, for this type of graphical model, the larger W_D the greater the likelihood of duplicates. We refer to W_G as the weight of the graph G, an unknown value, and to W_D as the sample weight, an observable quantity. As before, let D be a sample of order d. Evidently, high sample weight will support a hypothesis that sample contains duplicates, while low weight would not.

For a sample $D \subset V, |D| = k$, barring any additional information, the 'expected' weight of D is $W_G \left(\frac{\binom{k}{2}}{\binom{n}{2}} \right)$, but the value of W_G is unknown, it must be estimated from the sample. For samples $D_1, \ldots, D_{\binom{n}{k}}$, each of order k, consider the mean weight

$$\overline{W}_k = \frac{W_{D_1}, \ldots, W_{D_{\binom{n}{k}}}}{\binom{n}{k}}.$$

Then it's reasonable to estimate W_G with

$$\overline{W}_k \left(\frac{\binom{k}{2}}{\binom{n}{2}} \right) = \overline{W}_k \left(\frac{n(n1)}{k(k1-)} \right) \approx \overline{W}_k (\frac{n}{k})^2.$$

Note that if $k = n$, then $\overline{W}_k = W_G$, and if $k \approx \frac{n}{2}$, then $\overline{W}_k \frac{2n}{n}^2 = 4\overline{W}_{n/2}$, i.e.,$\overline{W}_{n/2}$ will under-estimate by a factor of 4.

The problem of identifying multiplicities (duplicates, triplicates, etc.) in a database is further complicated because records which presumably represent the same underlying entity aren't syntactically identical. Artificial intelligence and *fuzzy matching* algorithms are used to assign a numerical value (probabilities) which represent our belief that two records represent the same entity [3]. The algorithm (Database-to-graph Algorithm 1) below illustrates the essential features for the fuzzy matching Algorithm 1.

Algorithm 1 Database-to-graph Algorithm.

INPUT: N records $d_1, d_2, \ldots, d_N$ where each record is n−dimensional vector of attributes.
OUTPUT: Weighted graph (G, W).
1: **Step 1: Preprocessing.** Transform the N records into "standard" format. This step may involve various parsing and "data scrubbing"{The outputs of this step are vectors (processed records) $d'_1, d'_2, \ldots, d'_N$ where the dimensionality of $d'_k \geq n$.}
2: **Assign Fuzzy Matching.**Rename the outputs of Step 1, $x_1, x_2, \ldots, x_N$ & from now on refer to those vectors as nodes.
3: **foreach** pair $(x_i, x_j), i < j$ **do**
4: Calculate the weight of $w'_{i,j}$ of (x_i, x_j).
5: Normalize the weight $w_{i,j}$ of (x_i, x_j) so that $w_{i,j}$ are probabilities value, $0 \leq w_{i,j} \leq 1$.
6: **end foreach**

Notes – Algorithm 1.

1) This is a "one pass" algorithm that will terminate after $\binom{n}{2}$ steps.
2) We will not dwell on data structure or specific algorithm design issue except to say that once the algorithm is laid out, the rest is a standard programming task.
3) There are many ways to calculate weights, but they all are based on heuristics like *edit* or *Levenshtein* distance, Hamming distance, parametric statistical model, Bayesian models, or some subjective belief of the strength of association between pairs x_i and x_j.
4) Weight normalization should be done consistently as not to distort the original data.

This results in a graph where the weight $w(x, y)$ is a value between 0 and 1, with 0 signifies absence of an (undirected) edge between x and y, while 1 represents a certainty that x and y are connected. A value like $\frac{1}{2}$ would represent a complete uncertainty whether x is connected to y. In many cases this model is the only realistic approach to disambiguating entities. Therefore, weighted graph will mirror a database in the sense that unless two records are completely identical, the weight between their corresponding vertices will be strictly less than 1. We ignore 0 as non-edges. In applications we may think of an edge whose weight exceed some value as certain, so that we may decide that weight exceeding the value $p_{x,y} = 0.80$ as certain edge, while those less that $p_{u,v} < 0.80$ as fuzzy edges. The problem therefore is to infer as much as possible information about the graph, like number of edges (or those vertices with every pair connected with edges whose weights exceeding some threshold value 0.80), it's overall weight, the number of cliques, etc. The result is a pathway for further investigation and potential answers which, in turn may find 'many to many' or few issues. The result overall is a contribution to due diligence.

III. Conclusion

We suggest that this mathematical methodology can 'value add' to mechanisms for analysis to fulfill corporate responsibility in the context of business and human rights. The possibilities of our graph theory random sampling technique gives real potential for greater understanding of the networks in identity records and as a useful tool to identify clusters worthy of further examination.

IV. Acknowledgements

Peter Shaw is the recipient of the Research Start-up Fund of University of Information Science and Technology.

References

[1] "US Business Network for corporate responsibility news 1st October 2014," http://www.csreurope.org/us-develop-national-action-plan-consistent-un-guiding-principles-\business-human-rights, 2014, accessed: 2019-10-20.
[2] W. E. Winkler, "Record linkage," in *Handbook of statistics*. Elsevier, 2009, vol. 29, pp. 351–380.
[3] J. Barr and P. Shaw, "Ai application to data analysis, automatic file processing," in *2018 First International Conference on Artificial Intelligence for Industries (AI4I)*, 2018, pp. 100–105.

Authorized licensed use limited to: UNIVERSIDADE DE BRASILIA. Downloaded on November 05,2024 at 19:51:50 UTC from IEEE Xplore. Restrictions apply.