An Influence Assessment Method based on Co-Occurrence for Topologically Reduced Big Data Sets

Marcello Trovati and Nik Bessis
Department of Computing and Mathematics
University of Derby, United Kingdom
M.Trovati@derby.ac.uk
N.Bessis@derby.ac.uk

***Abstract*—The extraction of meaningful, accurate, and relevant information is at the core of Big Data research. Furthermore, the ability to obtain an insight is essential in any decision making process, even though the diverse and complex nature of big data-sets raises a multitude of challenges.**

In this paper, we propose a novel method to address the automated assessment of influence among concepts in big data sets. This is carried out by investigating their mutual co-occurrence, which is determined via topologically reducing the corresponding network. The main motivation is to provide a toolbox to classify and analyse influence properties, which can be used to investigate their dynamical and statistical behaviour, which would potentially lead to a better understanding and prediction of the properties of the system(s) they model.

An evaluation was carried out on two real-world data-sets, which were analysed to test the capabilities of our system. The results show the potential of our approach, indicating both accuracy and efficiency.

***Keywords*-Knowledge discovery; Large-scale Networks; Information extraction; Data analytics**

I. Introduction

Big Data is increasingly influencing the way we obtain, assess, and manage information. In particular, a very powerful and efficient approach to obtain an insight into real-world Big Data, is by determining the main properties that characterise such data-sets, and the factors that influence them.

There is extensive research on the automated extraction of causal relations between semantic objects, such as events, people, entities, factual data, etc. [4]. Evidence of causality is used in a variety of contexts, such as Bayesian networks (BNs), and decision trees. However, causality is a very strong statement regarding two or more concepts, which implies a direction ("A causes B" is not the same as "B causes A"), a general agreement of the definition of causality, e.g. what specific terms indicate a causal relationship, and a lack of ambiguity, especially in the automated extraction from textual sources. Although a detailed discussion of all the aspects related to causality goes beyond the scope of this paper (refer to [19] for an overview), we assume that *influence* describes a somehow weaker, yet more general concept of causality, at least from a semantic point of view. More specifically, influence between two or more objects may not be tied to a direction, or a well-defined, unambiguous semantic definition. Rather than giving its definition in terms of its specific semantic attributes, we regard the influence between two objects as a type of relation based on of their mutual co-occurrence. Clearly, co-occurrence does not imply any influence, since their mutual existence might be completely unrelated. However, when data-sets are modelled by networks with specific properties, such as scale-free, nodes refer to data that co-occur according to the types of connections defined by mutual edges. In particular, this kind of co-occurrence is likely to be associated with influence. In other words, when analysing data, if we can determine the co-occurrence of two or more elements, we can extract some intelligence which could be of significant importance. As introduced in [23], big data-sets can be topologically reduced to determine whether the data follow a scale-free or a purely random structure. This allows to classify the connections among elements in the data-set according to the properties associated with the corresponding network topology. In particular, if a data-set is best approximated by a random network then the existence of the edges follows a purely random distribution. In other words, it is assumed that there is no influence between the elements of the data-set. In fact, if a set of events is indeed random, no determined links between its elements can be established.

In this paper, we introduce a novel method to assess influence among concepts, based on co-occurrence, in big data-sets by approximating their structures with a topologically reduced network extracted from such data-sets. More specifically, given a real network, we provide a method to assess whether it can be approximated as either a random, or scale-free network to allow a better understanding of the associated data. In fact, when a data-set is shown to follow a scale-free topology, then the different elements exhibit well-defined relations. Furthermore, they also have the *preferential attachment* property [3]. Broadly speaking, this is equivalent to the concept that highly connected nodes are more likely to have new connections than less connected nodes. This property will be exploited in the method proposed in this paper, when the dynamical properties of topologically reduced networks are investigated. Subsequently, the level of influence between any two nodes is assessed, via the introduction of a suitably defined measure.

The main motivation of this paper is twofold. We first determined whether a real-world network can be modelled by a

specific network with reduced topology, as either a random or scale-free network. If the latter is the best fit, we then propose a method to assess the likelihood of an influence relation between two events based on their co-occurrence. Furthermore, our investigation does not only focus on structured data-sets, as we provide a text analysis capability to analyse textual information contained in the dataset(s). Such information is then converted into nodes in the extracted network to provide a full analysis of the relevant information [23].
The paper is structured as follows: in sections II and III an overview of the necessary theoretical background is discussed. Section V-A focuses on the main contribution discussed in this paper, namely influence measure and assessment. Section VI introduces the dynamical properties associated with a real-world Big Data system. Finally, sections VII and VIII address the evaluation and interpretation of the results, as well as future directions.

II. Big Data

Data is created around us at an increasing pace, raising complex challenges and crucial opportunities in the way we extract, assess and manage information. The ability to provide cutting-edge methods and tools to address the complexity posed by Big Data is extremely valuable, especially when applied to multi-disciplinary contexts [2].
Big Data is defined by the following properties, also called the *4 V's* [13]:

- **Volume:** the amount of data that is daily produced is impressive. The combination of real-time data with historical ones, can provide a crucial insight into the most appropriate and best decision process.
- **Velocity:** the information flow is continuously changing and pouring from a variety of data sources. The more of it that can be processed and assessed within specific time constraints, the better intelligence can be provided. However, this is raises a multitude of challenges, especially when combined with a large volume of data.
- **Variety:** data take a variety of shapes and forms. Streams of information can be collected from audio or video sources, as well as from sensors and textual sources, to name but a few. This diversity requires suitable tools and techniques that can be applied to efficiently deal with the different data types.
- **Veracity:** data contain erroneous, contradictory and missing information which potentially undermine the whole process of acquisition, assessment, and management of information. Therefore, tackling such issue is at the very core of Big Data science.

The method introduced in this paper addresses the above challenges specifically focusing on *volume* and *velocity* of Big Data. Although a formal evaluation is not carried out, we have noticed that the algorithms that define our approach, exhibit speed, efficiency, and computational power.
A crucial assumption in this paper is that such data-sets are based on one or more relation structures. In other words, it is possible to identify and classify the different data elements according to some semantic relationships. As an example, one of the data-sets that are considered in this paper, and described in section VII, contains unstructured data. However, its overall structure is determined by specific parameters, which allow the identification of some specific relations, such as geographical, and temporal.

III. Network Theory

Network theory has increasingly attracted much interest from a variety of interdisciplinary research fields, including mathematics, computer science, biology, and the social sciences. Their simple, yet effective formulation has allowed a successful exploitation of their applications in a wide range of real-world complex settings [27].
Formally, networks consist of a collection of nodes, called the node set $V = \{v_i\}_{i=1}^{n}$, which are connected as specified by the edge set $E = \{e_{ij}\}_{i\neq j=1}^{n}$ [3], excluding self-loops. We say that there is a path $p(v_i, v_j)$ between the nodes v_i and v_j, if we have a sequence of edges which connect a sequence of distinct nodes, such that it starts from v_i and ends at v_j.
In this paper, we focus on a specific use of data and text mining techniques to determine the best topological reduction that approximates a real-world data set, based on the assumption that it follows the "4 V's" characterisations. As mentioned above, this is based on research carried out in [23], which has been subsequently improved and expanded, especially the text and data mining capabilities. In fact, as described in section IV, the text mining techniques proposed in [23] have been further developed to provide a wider range of more accurate results via the extraction of nodes from textual sources.
The topological properties of the network is then fully analysed to determine the path-connections, which provide an insight into the co-occurrence, and subsequently the mutual influence of any two concepts corresponding to specific nodes.
In the rest of this section, we give an overview of random and scale-free networks, which are used to topologically reduce real-world networks.

A. Random Networks

Random graphs are defined by a random process which governs their overall topology, as the existence of edges between nodes depends on a probability p. Such networks have been extensively investigated since the dawn of Graph (and Network) theory, and a variety of properties have been identified depending on their theoretical, or applied context. In particular, the fraction p_k of nodes with degree k follows

$$p_k \approx \frac{z^k e^{-z}}{k!},$$

where $z = (n-1)p$ [5].
One of the crucial aspects of a random network is the fact that the set of relationships among the concepts modelled by the edges and nodes respectively, are purely random.
As mentioned above, when the nodes are associated with events, or semantic objects, from a data-set containing information semantically linked, we can interpret the edges as a relationship between the nodes they connect. Therefore, if a network is topologically reduced to a purely random structure,

we will assume that the relationships captured by the edges do not indicate a co-occurrence, and more specifically, influence. This is due to the fact that a random network is associated to a purely randomised system, and the co-occurrence of nodes does not follow a determined law. This fact will be exploited in our method as described in the next sections.

B. Scale-Free Networks

Scale-free networks appear in a multitude of contexts, including the World Wide Web links, as well as biological and social networks [3]. Furthermore, the continuous improvement of data extraction tools is leading to the identification of more instances of such networks.

The main property of scale-free networks is related to their node degrees which are governed by a power law. More specifically, for large values of k, the fraction p_k of nodes in the network having degree k, is modelled as

$$p_k \approx k^{-\gamma} \tag{1}$$

where γ has been empirically shown to be typically in the range $2 < \gamma < 3$ [3].

A direct consequence of Equation 1, is that there is a relatively high likelihood to have hubs, which characterise the topological properties of the corresponding networks, as well as the way information spreads across them [9], [14].

As discussed above, an important feature of such networks refers to the fact that new nodes are created over time, which are likely to be connected to existing nodes that are already well connected. This principle of "preferential attachment" will be discussed in section VI. Furthermore, since the connectivity of nodes follows a distribution which is not purely random, we make the assumption that networks that are topologically reduced to scale-free structures, the edges indicate an co-occurrence-based influence between the corresponding nodes. More specifically, the dynamical properties of such networks provides an insight into their evolving which can lead to predictive capabilities.

C. Reduced Network Topology Extraction

Big data-sets can be efficiently analysed by reducing their topology [23], [25]. In other words, they can be automatically "approximated" as either random or scale-free networks. The importance of such process is twofold. Firstly, it allows the identification of a topological structure which can give an insight into the corresponding data-sets. Secondly, provided the assumption of co-occurrence-based influence discussed in section II, we can extract information on the system modelled by such network that can be used to determine relevant intelligence.

As discussed in section VII, his method has shown to produce relevant and accurate results, with the important feature of being computationally scalable and optimised to address Big Data issues. Furthermore, a set of text mining techniques is also integrated to provide analytical capabilities to analyse both structured and unstructured data-sets. In [23], the algorithms utilised for the reduced network topology extraction process are introduced, and the reader can refer to that article for further details. Furthermore, these algorithms also allow the identification of the long-tail distribution in the case of scale-free networks, resulting in a more accurate and relevant extraction [25].

IV. Information Extraction Via Text Mining Techniques

Due to the diverse nature of Big Data, it is essential to provide information extraction capabilities from unstructured data. In this paper, we consider text mining techniques to allow the identification and assessment of relevant information from textual data sources [7].

Depending of the general context and the given semantic information, a variety of text mining techniques can be used, which in general depend on the type of data and their structure. In particular, sentiment analysis [15], focuses on the detection of "opinions" or *polarity* from textual data sources. The method introduced in [23] specifically targeted a data-set containing information on air accidents and near misses [1]. More specifically, some of the entries consisted of pilots comments.

In this paper, we have expanded this method by, first of all, improving the vocabulary containing the keywords as in [23]. These included a list of words suitably describing the associated polarity. An extensive set of new keywords and cue phrases was created by automatically extracting them from the tagged version of the Brown Corpus, which contains approximately 500 samples of English-language texts [12]. This was carried out by considering the triples `(NP1, VP, NP2)` where

- `NP1` and `NP2` are the *noun phrases*, i.e. phrases with a noun as its head word [11], which had to contain one or more keywords from [23]. Note this requirement had to be satisfied for *at least one* of the `NP`s, and not just for both of them.
- `VB` is the *linking verb*.

Subsequently, the extracted `NP1` and `NP2` were manually analysed to identify the appropriate keywords, and cue phrases. A detailed evaluation of this approach goes beyond the scope of this paper, since it specifically addresses issues that are not directly relevant in this context. However, we tested it on two randomly chosen papers [20], [21]. The automatic extraction was compared with a manual one, which produced a recall of 65% and a precision of 74%.

Similarly to [23], the following steps were included:

- Textual fragments from input data-sets were first shallow parsed via the Stanford Parser [7].
- A grammar-based extraction identified triples of the form `(NP, verb, keyword)`, where `NP`, is the noun phrase, `verb` is the linking verb, and `keyword` consists of one or more keywords as mentioned above.

The triples are used to populate the nodes and edges of the corresponding network, by identifying any connection among the keywords defined above, with the corresponding elements of the data-sets. In order to avoid any redundancy, all the extracted terms were normalised, where *normalisation* is the

process of mapping different variants of a term to a unique and standardised form [11]. For example, an entry such as

> *"A P180 was 10 left for traffic. CSV HI called and gave me control of the P180 to turn him back on course. As I cleared him to his destination; UES; I said 'cleared direct Waukesha' He hesitated and then said 'where.' Then I said 'cleared direct to your destination' He said oh; Ok. My D-Side and I were still a little uneasy so I went back and asked to verify his destination; which he said; was 'MKC.' Atlanta Center must have changed the destination in his routing. First of all; don't say 'cleared direct destination.' Be specific with the name and/or identifier of the airport. Second of all; be careful when using a splat or down arrow. I don't see why this was even used for him; but I don't know the situation in ZTL."*

would produce the output in table I

TABLE I
THE RESULT PRODUCED WITH THE METHOD DESCRIBED IN SECTION III-C

Term		Normalisation
hesitated	→	hesitation
uneasy	→	uneasiness
don't know	→	uncertainty
be careful	→	lack of carefulness
be specific	→	lack of specificity

In the above example, the term on the right hand side column would define connected individual nodes as part of the corresponding network.

V. Description of the Method

In this section, we describe the method by first introducing the relevant mathematical background. This will be exploited in section V-C, where the main algorithms are defined and discussed.

A. Measuring Co-Occurrence-based Influence

Relation discovery is essential in assessing and predicting how knowledge spreads and evolves. In [28], an automated construction and annotation of biological networks is investigated, and an influence measure based on co-occurrence is introduced. This is based on co-occurrence of specific (linguistic) terms to establish semantic and linguistic relationships between them. Furthermore, the frequency of this type of co-occurrence usually follows a scale-free distribution, which is then investigated to determine fuzzy-sets membership. The method proposed in this paper follows, broadly speaking, the opposite direction. In fact, we start from a real-world network, provided it follows a scale-free network, and assess a co-occurrence measure based on the topological properties of such network.

In [22], a method to evaluate the influence and direction between two concepts in a semantic network is discussed. In order to achieve this, the authors introduce a scalable approach to assess the relevant parameters which determine the way one semantic entity influences another one. In particular, the dynamics of such parameters is a crucial aspect which models the overall properties of information propagation and its assessment. Despite in this paper we have not addressed semantic networks in general, provided that a data-set can be topologically reduced to a scale-free network, or in other words it is not purely random, we must clarify some important points.

First of all, and probably most importantly, the fact that two nodes in the reduced topology network lie on the same path, does not imply any influence. It is rather an indication of co-occurrence, that is the two events, or concepts, associated with those two nodes, may occur together depending on the properties of the paths joining them. Assuming this entails an influence relation, or even a causal relation, would clearly be an erroneous supposition. For example, if whenever it rains I open my umbrella, I can by no means assume that raining causes my umbrella to open. However, ascertaining co-occurrence is a fundamental step to determine, and often super-impose, relation networks defined by the nodes of a general network. These are, as the name suggest, defined by nodes which are connected by edges whenever a relation exists between them. Due to their nature, relational networks include a large variety of examples, from general semantic networks to causal networks. In particular, such networks can provide the ability to extract intelligence from a data-set, which leads to a better and more comprehensive understanding of the information related to it. In fact, understanding and assessing the interconnection among data provide useful modelling frameworks which can be applied to risk and decision analysis [26].

B. Mathematical Formalism

In this section, we introduce the mathematical formalism used to measure and assess the co-occurrence-based influence of two events. Furthermore, this will also show the main computational issues and give an insight into our motivation to provide a better approach.

Let $p_{k,l}(v_{i_1}, v_{i_l})$ be the weighted probability of choosing the $k-$path of length l between two nodes v_{i_1} and v_{i_l}. In other words,

$$p_{k,l}(v_{i_1}, v_{i_l}) = \frac{w_{(v_{i_1}, v_{i_2})}}{\deg(v_{i_1})} \prod_{j=2}^{l-1} \frac{w_{(v_{i_j}, v_{i_{j+1}})}}{\deg(v_{i_j}) - 1}, \tag{2}$$

where $\deg(v)$ is the degree of the node v, and $w_{(v_{i_a}, v_{i_{a+1}})} \in (0, 1]$ is the weight of the edge joining v_{i_a} and $v_{i_{a+1}}$. Therefore, the (weighted) probability $\Delta_l(v_{i_1}, v_{i_l})$ of choosing a path of length l between the two nodes (not necessarily edge-independent) is

$$\Delta_l(v_{i_1}, v_{i_l}) = \sum_{k=1}^{N_l(v_{i_1}, v_{i_l})} p_{k,l}(v_{i_1}, v_{i_l}), \tag{3}$$

where $N_l(v_{i_1}, v_{i_l})$ is the number of paths of length l. The weight of the edges depends on a variety of parameters, which

are established when defining the network.
However, one of the main drawbacks of Equations 2 and 3 is that their direct implementation can be quite expensive from a computational point of view, and most of the existing algorithms are based on the breadth-first search whose time complexity is $O(|V| + |E|)$ [6]. As a consequence, we are proposing a method to assess co-occurrence between two nodes based on the following aspects:

- The shortest path(s) between two nodes and their number,
- The degree of the nodes along those paths

More specifically, the first point gives crucial information on the connected-ness of the two nodes, whilst the last one provides information on the topological properties of such paths. Furthermore, since many real-world networks are based on big data-sets, there are clearly important issues in terms of computational efficiency that need addressing. As a consequence, we propose a method which is applicable in a big data scenario, due to its computational efficiency, accuracy and scalability.

C. Description of the Algorithm

As mentioned above, the above equations raise computational challenges that are often too complex to fully address, especially when dealing with large data-set which may vary in real time. The algorithm we are proposing, is specifically designed to deal with this scenario, providing an accurate, agile and scalable approach.
Rather than finding *all* the paths between two nodes, which would be then compared with all the other ones in the network, we only focus on the shortest path(s) between them. In fact the computation of the shortest paths is based on much more efficient algorithms [6]. Furthermore, we only consider the local properties of the nodes, so that we will not consider the overall edges in the network (which is typically very large). An important property that we will exploit is that *independent-edge* paths show a stronger co-occurrence-based influence relationship than paths which share common edges. Menger's theorem [16] describes the fact that the maximum number of pairwise edge-independent paths between two nodes is the same as the size of the minimum edge cut for those two nodes. However, the latter – also know as the min-max cut flow problem – may not be solvable for general networks [5]. With this in mind, we propose the following algorithm

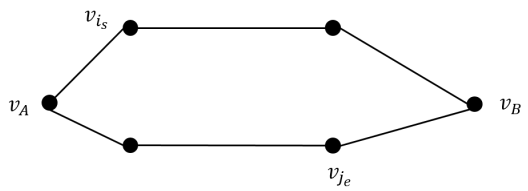

Fig. 1. Depiction of Algorithm 1 described in section V-C. Here, the shortest path $p(v_{i_s}, v_{i_e})$ has length 3, which is also the length of the path $p(v_A, v_B)$.

This algorithm has shown to provide an efficient alternative to the min-max flow problem for this particular context.
Similarly to Equations 2 and 3, an important property we have to consider is the degree of the nodes along each path.

Algorithm 1: Relative strength between v_A and v_B.

Data: Two nodes v_A and v_B
Result: Relative strength $\tilde{L}(v_A, v_B)$ between v_A and v_B

Obtain all the shortest paths between two nodes, say v_A and v_B, and call them $p_1(v_A, v_B), \dots p_n(v_A, v_B)$, with length $l(p_i(v_A, v_B)) = n$, for any $1 \leq i \leq n$;
for *All the couples* $(p_i(v_A, v_B), p_j(v_A, v_B))_{i \neq j=1}^{n}$ *consisting of the* $n!$ *permutations of the different paths.* **do**
  Choose v_{i_s} and v_{j_e} ("s" stands for *start*, and "e" for *end*) so that the former is the first node on p_i *after* v_A and the latter is the node on $p_j(v_A, v_B)$ *before* v_B as depicted in Figure 1;
  Find the shortest path(s) between v_{i_s} and v_{j_e}
  if *it has length equal to* $l(p_i(v_A, v_B))(= l(p_j(v_A, v_B)))$ **then**
    assume that $L_k(v_A, v_B) = 1$;
  else
    $L_k(v_A, v_B) = 0.5$;
  end
end
Evaluate $\tilde{L}(v_A, v_B) = \text{average}(L_k(v_A, v_B))$.

Loosely speaking, we still want to determine the weighted probability of reaching a specific node from another one. As a consequence, from (2) we define

$$W_{(v_{i_1}, v_{i_2})} = \frac{w_{(v_{i_1}, v_{i_2})}}{\deg(v_{i_1})}, \tag{4}$$

where $0 \leq w_{(v_{i_1}, v_{i_2})} \leq 1$. The weights $w_{(v_1, v_2)}$ between any two nodes v_1 and v_2, can be specified prior to the implementation of the different algorithms, or via machine learning techniques, which can be integrated into the process. In this paper, we assume that the weights w's between nodes are all equal to 1.
We then define the *co-occurrence-based influence measure* $\Lambda(v_{i_1}, v_{i_l})$ between the nodes v_{i_1} and v_{i_l} as

$$\Lambda(v_{i_1}, v_{i_l}) = \frac{1}{(|P(v_{i_1}, v_{i_l})| - 1) \sum_{P_i \in P(v_{i_1}, v_{i_l})} W_{P_i}} \left[\tilde{L}((v_{i_1}, v_{i_l})) \sum_{i \neq j=1}^{|P(v_{i_1}, v_{i_l})|} \left(\prod_{i=2}^{l-2} W_{(v_i, v_{i+1})} + \prod_{j=2}^{l-2} W_{(v_j, v_{j+1})} \right) \right]. \tag{5}$$

where $W_{P_i} = \frac{w_{(v_{i_1}, v_{i_2})}}{\deg(v_{i_1})}$, and $P(v_{i_1}, v_{i_l})$ is the set of shortest paths between v_{i_1} and v_{i_l}. Also note that

$$0 \leq \Lambda(v_{i_1}, v_{i_l}) \leq 1. \tag{6}$$

The co-occurrence-based influence measure $\Lambda(v_{i_1}, v_{i_l})$ not only does give an insight on how likely two concepts are to co-occur, but it also provides an evaluation of the way two nodes are connected in a network. In algorithm 1, there are

many parameters whose values depend on the network which is being considered, and more importantly, how they were described. Therefore, $\Lambda(v_{i_1}, v_{i_l})$ provides values that need to be interpreted, either automatically or manually, to determine the likelihood of co-occurrence. In this paper, we assume that $\Lambda(v_{i_1}, v_{i_l}) \geq 0.5$ indicates a high level of co-occurrence-based influence. This has been determined by manually assessing the co-occurrence-based influence measure for different nodes. More specifically, we have algorithm 2:

Algorithm 2: Algorithm measuring the co-occurrence between $\Lambda(v_A$ and $v_B)$

Data: Two nodes v_A and v_B
$\tilde{L}(v_A, v_B)$
Result: Co-occurrence-based influence measure between two nodes, say v_A and v_B

Obtain $P(v_A, v_B)$ the set of all the shortest paths between v_A and v_B;

Evaluate $\Lambda(v_A, v_B)$, using Equation 5.

VI. The Dynamics of Co-occurrence

As discussed earlier, the fact that an edge between two nodes refers to their mutual co-occurrence, can be exploited to provide some predictive power in terms of the dynamical properties of the associated network. In fact, the properties of preferential attachment for scale-free networks can give a useful insight into the dynamics of the overall network when the creation of new nodes and/or new edges is investigated [3]. Since co-occurrence-based influence is assessed according to the method introduced above, we can apply equation 5 when the shortest paths between two nodes are modified by creating new nodes and/or adding more connecting edges. In particular, the concept of *velocity* in Big Data can be regarded as the addition and/or removal of either edges or nodes in the corresponding network. As a consequence, in this paper we consider the following two scenario:

A) A new edge is introduced at each "time" interaction, and no new nodes are created;
B) A new node is introduced, corresponding to a new event, which has to be investigated.

These scenarios can be re-phrased, and simplified, as follows: assume we have a set of shortest paths between two nodes v_A, and v_B. What happens to $\Lambda(v_A, v_B)$ when a new node and/or new edges are introduced?
Note that we have the following cases

- An extra edge between two nodes on the shortest paths between v_A, and v_B, is added. This will create an extra path $p(v_A, v_B)_N$.
  - If $l(p(v_A, v_B)_N)$ is less or equal to the shortest path length, then

Algorithm 3 provides a sequence of $\Lambda(v_A, v_B)_t$, for $t = 0, \ldots$, so that

1) If $\Lambda(v_A, v_B)_t$ tends to a specific (finite) value, we have *stability*, and

Algorithm 3: Algorithm assessing the co-occurrence-based influence measure.

Data: Two nodes v_A and v_B
Result: Sequence of $\Lambda(v_A, v_B)_t$, for $t = 0, \ldots$ between v_A and v_B

Obtain $P(v_A$ and $v_B)$, i.e. the set of all the shortest paths between v_A and v_B ;

if *An extra edge between two nodes on two separate shortest paths between v_A, and v_B, is added. This will create an extra path $p(v_A, v_B)_N$* **then**
  if *$l(p(v_A, v_B)_N)$ is less or equal to the shortest path length* **then**
    update $\Lambda(v_A, v_B)_{t+1}$;
  else
    $\Lambda(v_A, v_B)_t = \Lambda(v_A$ and $v_B)_{t+1}$;
  end
end
if *One of the nodes, v_h, on a shortest path between v_A, and v_B, has an extra edge connected to another node not part of any shortest paths* **then**
  Update $\Lambda(v_A, v_B)_{t+1}$ with new value of degree;
end

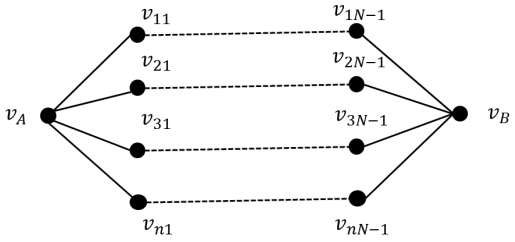

Fig. 2. Example of scenario A, when $N = 4$.

2) If $\Lambda(v_A, v_B)_t$ oscillates within a range, we might not have stability. In fact, this would depend on the oscillation range, and in what case this would be considered too big.

A full investigation of the dynamical properties goes beyond the scope of this paper, and will be the focus of future research.

VII. Evaluation

In this section, we discuss our evaluation of our approach which is based on the data sets [10], and [1], which were investigated in [23] and [24]. The first data-set has structured data types, and in this case no text mining method was applicable. We had a generated a network with $3,046$ nodes and $11,794$ edges. As described above, we implemented our algorithm to reduce the topology of the networks associated with the given dataset. More specifically, we considered specific parameters, namely the date of the earthquake activity, its geographical location, time of the day, and its intensity, and assessed their corresponding reduced-topology networks, see [24] for more details. This produced, using the notation introduced in section III-B, the results shown in table II.

TABLE II
THE RESULTS OF THE REDUCED-TOPOLOGY ALGORITHM, AS IN [24]

Type of Seismic Event	γ **Values**
Date of seismic activity	2.17
Geographical Location of Seismic Activity	2.74
Time of Seismic Activity	2.93
Intensity of Seismic Activity	2.89

The second data-set contains information regarding air accidents and near-misses, which we have been re-valuated with the improved text mining method discussed in section IV. Notably, an increased number to nodes were extracted, due to the better extraction capabilities. In particular, the generated network had $3,237$ nodes and $12,803$ edges. However, this change in the number of nodes and edges was not reflected in the reduced topology, which is still best approximated by a scale-free network with parameters $\gamma = 2.04$, and $\sigma = 0.34$. Note that this was evaluated using the algorithm described in [25], which addresses the long-tail distribution in scale-free networks. Note that γ is in interval between 2 and 3, which is consistent with the experimental findings in [3].
The data-set [10] was also used to investigate the method proposed in section V, was carried out automatically and subsequently compared with a manual extraction. However, due to the size of the network, we only concentrated on a smaller sub-network, containing approximately 400 nodes, to facilitate the task. Note that the data-set contains entries related to the time of seismic activity. When populating the corresponding network, these were grouped into the same node if they were within 10 minutes apart. Clearly, this type of grouping is open to interpretation as we will discuss shortly.
Three particular scenarios were considered focusing on the geographical location of seismic activity. More specifically, we obtained the results as described in table III.

TABLE III
THE EVALUATION OF Λ BY USING ALGORITHM 3

Node 1	**Node 2**	Λ
Adriatic (area)	Southern Italy	0.77
Aegean Sea	Central Italy	0.54
Estonia	Crete	0.39

This was subsequently assessed manually to determine whether these results were relevant and accurate. Interestingly, the only instance that caused some disagreement between the automatic and manual evaluation related the two nodes "Aegean Sea" and "Central Italy". In fact, it was suggested that there were indeed a co-occurrence-based influence between them. However, the relatively low value of $\Lambda = 0.54$, was due to the time entries. If nodes were created by merging times within 30 minute slots, Λ increased to 0.63, suggesting a stronger level of co-occurrence between the two nodes.
Intuitively, the above results appear consistent as one would expect some kind of influence of seismic activity between geographical locations relatively close to each other.

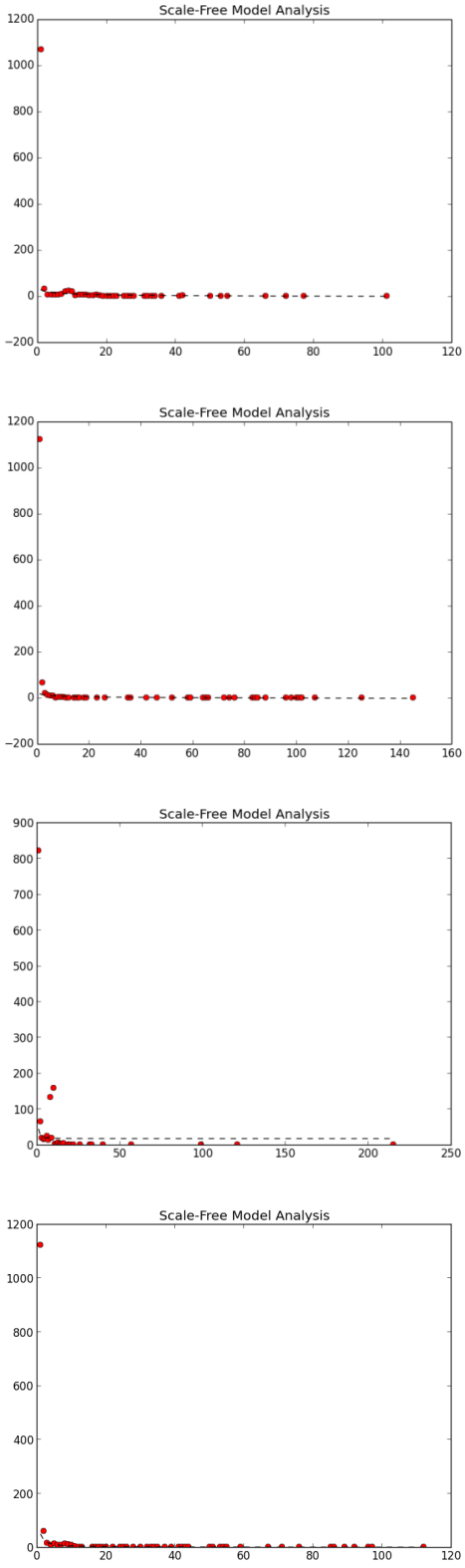

Fig. 3. The Scale-Free Network Structure of the data-set introduced in [23], focusing of the geographical, date, hour and intensity of the seismic activity, respectively.

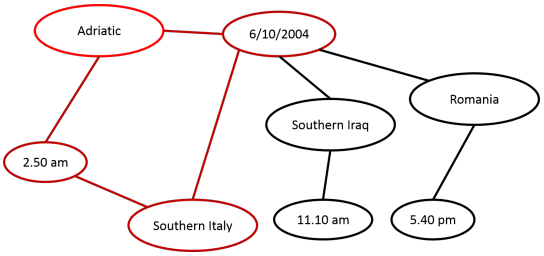

Fig. 4. Example of a small section of the sub-network around the nodes"Adriatic" and "Southern Italy", as discussed in section VII. Two shortest paths connecting them are highlighted in red.

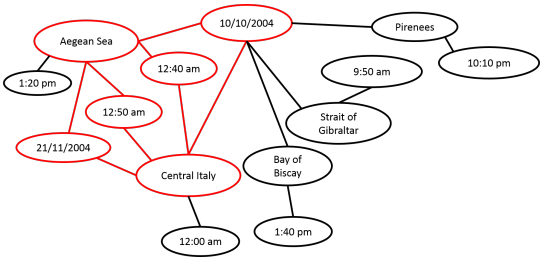

Fig. 5. Another example of a small sub-network, which is centred at nodes "Aegean Sea", and "Central Italy". Note the three shortest paths which are highlighted in red.

VIII. Discussion and Future Work

The use of co-occurrence as an indication of influence, has many potential applications. In fact the associated networks provide an insight into the relational structure of the corresponding data-set. Their extraction from data is typically a complex task, which greatly depends on its nature, what type of relations are to be modelled, and what purpose such representation needs to serve. A specific example is Bayesian Networks. These are acyclic networks, i.e. no loops are present, so that nodes represent random variables and edges represent conditional dependencies. Two nodes which are not connected by an edge correspond to variables that are conditionally independent of each other. Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and gives the probability of the variable represented by the node. The automatic extraction of BNs has been extensively investigated [18], based on textual properties which describe the types of relationships the concepts associated with nodes are linked by. However, defining and populating BNs is typically a complex task, due to their probabilistic and mathematical constraints. We believe that the method described in this paper, would provide an effective approach to this challenge. In fact, much research on BNs has focused on the analysis of causal relationships [18]. On the other hand, the conditional dependencies among variables can successfully described by influence relationships. However, this is beyond the scope of our paper, and more research is needed in order to achieve meaningful and complete BNs.

In this paper, we propose a novel method to assess the influence between nodes of topologically reduced networks extracted from data, by investigating their co-occurrence. The aim was to provide an efficient and accurate tool to generate intelligence from Big Data. The evaluations that we have carried out indicate the potential of our approach, as well as promising new research directions that will be pursued in future research. This will include the extraction of BNs as described in the previous section, which is part of a larger line of inquiry focusing on the creation of a toolbox to facilitate and guide the decision making process, supported by the identification of intelligence from Big Data.

References

[1] Aviation Safety Reporting System Database. Available from http://asrs.arc.nasa.gov/search/database.html. [1 March 2014]

[2] Azar, A T, and Hassanien, A E Dimensionality Reduction of Medical Big Data Using Neural-Fuzzy Classifier. *Soft Computing*, 1432–7643, 2014

[3] Albert R and Barabási A L Statistical Mechanics of Complex Networks. *Reviews of Modern Physics*, 74, 47, 2002

[4] Blanco E, Castell N, and Moldovan D Causal Relation Extraction. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008

[5] Bollobás B Modern Graph Theory. *Graduate Texts in Mathematics*, vol. 184, Springer, New York, 1998

[6] Cormen T H, Leiserson C E, and Rivest R L Introduction to Algorithms. *MIT press*, 1990

[7] De Marneffe M F, MacCartney B and Manning C D Generating Typed Dependency Parses from Phrase Structure Parses. *Proceedings of LREC*, 2006

[8] Diehl C P, Namata G, and Getoor L Relationship Identification for Social Network Discovery. *Proceedings of the 22Nd National Conference on Artificial Intelligence* Vol 1, 2007

[9] Ebel H, Mielsch L I and Bornholdt S Scale-free Topology of E-mail Networks. *Phys. Rev.* E 66, 035103, 2002

[10] European-Mediterranean Seismological Centre Database. Available from http://www.emsc-csem.org/. [1 May 2014]

[11] Feldman R, and Sanger J The Text Mining Handbook. *Cambridge University Press* 2006.

[12] Francis, W N, and Kucera, H The Brown Corpus: A Standard Corpus of Present-Day Edited American English. *Providence, RI: Department of Linguistics, Brown University* [producer and distributor], 1979

[13] GUPTA R, GUPTA H, AND MOHANIA M Cloud Computing and Big Data Analytics: What Is New from Databases Perspective? *Big Data Analytics*, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 42-61, 2012

[14] HEIN O, SCHWIND M AND KONIG W Scale-free networks. *WIRTSCHAFTSINFORMATIK* 48:4, 2006

[15] LIU B Sentiment Analysis and Opinion Mining. *Morgan and Claypool Publishers* 2012

[16] MENGER K Zur allgemeinen Kurventheorie. *Fund. Math.* 10: 96115, 1927

[17] NEWMAN M E J AND PARK J Why Social Networks Are Different From Other Types of Networks. *Physical Review E.* 68(3) 36122, 2003

[18] SANCHEZ-GRAILLET O AND POESIO M Acquiring Bayesian Networks from Text. *Proceedings of LREC*, European Language Resources Association, 2004

[19] SCHAFFER J Causation, Influence, and Effluence. *Analysis* 61: 1119, 2001. doi: 10.1111/1467-8284.00263

[20] SCHONBAUER R, SOMMERS P, MISFELD M, DINOV B, FIEDLER F, HUO Y, AND ARYA A Relevant ventricular septal defect caused by steam pop during ablation of premature ventricular contraction. *Circulation* 2013

[21] SOHEILYKHAH S, SHEIKHANI A, SHARIF A G, AND DAEVAEIHA M M Localization of Premature Ventricular Contraction Foci in Normal Individuals Based on Multichannel Electrocardiogram Signals Processing. *Springerplus*, 486, 2013

[22] TROVATI M AND BAGDASAR O Influence Discovery in Semantic Networks: An Initial Approach. *Proceedings of UKSim*, 2014

[23] TROVATI M, BESSIS N, HUBER A, ZELENKAUSKAITE A AND ASIMAKOPOULOU E Extraction, Identification and Ranking of Network Structures from Data Sets. *Proceedings of CISIS*, pp:331-337, 2014

[24] TROVATI M, ASIMAKOPOULOU E, AND BESSIS N An Analytical Tool to Map Big Data to Networks with Reduced Topologies . *Proceedings of InCoS*, pp: 411-414, 2014

[25] TROVATI M Reduced Topologically Real-World Networks: a Big-Data Approach *International Journal of Distributed Systems and Technologies*, In press 2015.

[26] ZELENKAUSKAITE A, BESSIS N, SOTIRIADIS S AND ASIMAKOPOULOU E Interconnectedness of Complex systems of Internet of Things through Social Network Analysis for Disaster Management *Proceedings of 4th IEEE INCoS-2012*, pp: 503-508

[27] WATTS D J AND STROGATZ H S Collective Dynamics of Small-World Networks. *Nature*, 393, pp. 440-442, 1998

[28] WREN J D Using Fuzzy Set Theory and Scale-free Network Properties to Relate MEDLINE Terms. *Soft Computing*, 10,4, 2006