# Automated Extraction of Fragments of Bayesian Networks from Textual Sources

Marcello Trovati

*Department of Computer Science, Edge Hill University, UK*

Jer Hayes

*IBM Research, Dublin Lab, Ireland*

Francesco Palmieri

*Department of Computer Science, University of Salerno, Italy*

Nik Bessis

*Department of Computer Science, Edge Hill University, UK*

**Abstract**

Mining large amounts of unstructured data for extracting meaningful, accurate, and actionable information, is at the core of a variety of research disciplines including computer science, mathematical and statistical modelling, as well as knowledge engineering. In particular, the ability to model complex scenarios based on unstructured datasets is an important step towards an integrated and accurate knowledge extraction approach. This would provide a significant insight in any decision making process driven by big data analysis activities. However, there are multiple challenges that need to be fully addressed in order to achieve this, especially when large and unstructured data sets are considered.

In this article we propose and analyse a novel method to extract and build fragments of Bayesian Networks (BNs) from unstructured large data sources. The results of our analysis show the potential of our approach, and highlight its accuracy and efficiency. More specifically, when compared with existing approaches, our method addresses specific challenges posed by the automated extraction of BNs with extensive applications to unstructured and highly dynamic data sources.

The aim of this work is to advance the current state-of-the-art approaches

to the automated extraction of BNs from unstructured datasets, which provide a versatile and powerful modelling framework to facilitate knowledge discovery in complex decision scenarios.

*Keywords:* Text Mining, Network Theory, Bayesian Networks

## 1. Introduction

Big Data has been drawing increasing attention from numerous research communities, which has led to the development and advancement in its theoretical foundations and applications to address the challenges raised by this field. Loosely speaking, Big Data is characterised by the *4 V*'s, namely volume, velocity, variety and veracity (Trovati M, 2015 A), and different approaches might address some of these aspects. In recent times, the analysis of unstructured data from textual sources has played a strategic role in many activities influencing our daily life. Such data are often characterised by a certain degree of uncertainty, as well as inaccuracies, which raise numerous challenges for the accurate extraction of new knowledge (AnHai D, 2006). Furthermore, the ability to identify useful connections between concepts, insights and trends from such analysis (in order to automatically build knowledge for supporting decision making processes) is at the heart of cutting-edge research in several research fields with a multitude of applications in many disciplines (Trovati M, 2015 A).

The extraction, classification and aggregation of probabilistic information is crucial in modelling complex systems, and the interconnections between the relevant concepts and associated variables underpin the general properties of such systems (Dojer N, 2013). In particular, their characteristics are shaped by statistical properties linked to the graph-theoretic features of the associated semantic networks. These can be modelled in terms of nodes, corresponding to specific words or concepts, together with their mutual connections, representing the semantic associations between them. Interestingly, many of such networks exhibit a small-world structure characterised by the combination of highly clustered neighbourhoods, a sparse connectivity and a short average path length (Watts DJ, 1998), as well as scale-free organisation (Albert R, 2002). This is characterised by a node degree distribution following a power law structure. Any knowledge process modelled by such networks reflects the semantic network growth dynamics. Each new node added to the semantic network immediately inherits some of the connections

characterising the pre-existing nodes (identified as its neighbourhood). Furthermore, highly connected nodes are more likely to be associated with more significant knowledge (Watts DJ, 1998). The structural principles characterising scale-free networks potentially have very important implications in understanding the development of new knowledge. In fact, this facilitates the abstract understanding of how semantic organisations can evolve and grow, based on basic statistical criteria. This enables knowledge construction in the form of probabilistic relationships between concepts (Trovati M, 2014).

Bayesian Networks (BNs) (Jensen FV, 2009), (Pearl J, 1998) are graphical structures with emphasis on cause and effect modelling in many knowledge-related domains (Ji J, 2011), (Cruz-Ramirez N, 2009). Their main characteristic is the ability of capturing the probabilistic relationship between variables, as well as their historical information. In particular, they facilitate the creation of systems by modelling knowledge in a way that is easily comprehensible by humans (Rance B, 2012), (Kuipers BJ, 1985), (Kuipers BJ, 1984). More formally, BNs are directed acyclic graphs where their nodes represent Bayesian random variables. In other words, they are associated with observable quantities, unknown parameters, hypotheses, etc. Nodes that are are conditionally dependent are joined by an edge, and each node is associated with a probability function whose input is a set of values from its parent nodes (i.e. the nodes connected to it). The output is the probability of the variable represented by the node.

BNs have proved to be very successful when a scenario consisting of pre-acknowledge information coupled with uncertain or partially known data, is considered (Pearl J, 1998). The extraction of BNs from text is typically a complex task due to the intrinsic ambiguity of natural language (Sanchez-Graillet O, 2004). In fact, BNs are defined by strict topological and probabilistic rules, which are difficult to fully automatise. Issues such as low recall and precision, as well as contradictory information, must be dealt with by any BN extraction tool. Therefore, they tend to rely on a substantial level of human supervision and interaction (Raghuram S, 2009).

In this article, we introduce a systematic method to extract and populate fragments of BNs (and hence structured knowledge) from textual sources, based on grammar and lexical properties, as well as on the topological features of networks subsequently being extracted. The aim is to provide an ag-

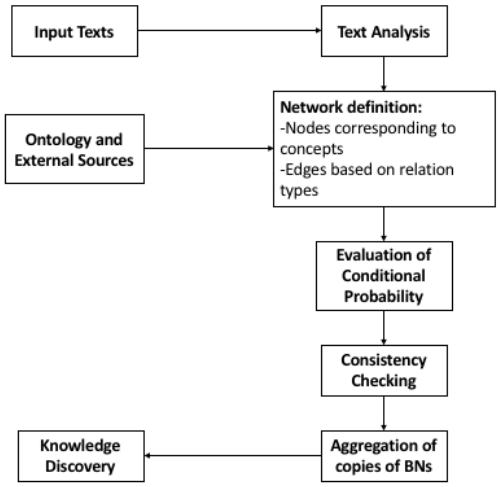



Figure 1: The general architecture of the scheme presented in Section 2.

ile yet accurate method to identify and assess probabilistic relations between concepts, which are subsequently embedded onto suitable BNs. However, this is typically a complex problem especially, when addressing the emergence of large volumes of unstructured data sets, typically referred to as Big Data. These are now considered to be one of the most promising sources for knowledge extraction.

The paper is structured as follows: in Section 2 we describe the main architecture of the proposed system; Sections 3 and 4 discuss the ontology used in this context, and the techniques and algorithms that can be used to extract probabilistic information. Sections 5 and 6 focus on the network topology, probabilistic information extraction, and knowledge discovery. In Section 7 we discuss the implementation and evaluation of our approach. Finally, Section 8 addresses future work and research directions.

## 2. The Architecture of the Extraction Scheme

As depicted in Figure 1, the knowledge extraction scheme, which finally results in the creation of suitable BNs, is structured in the following basic components:

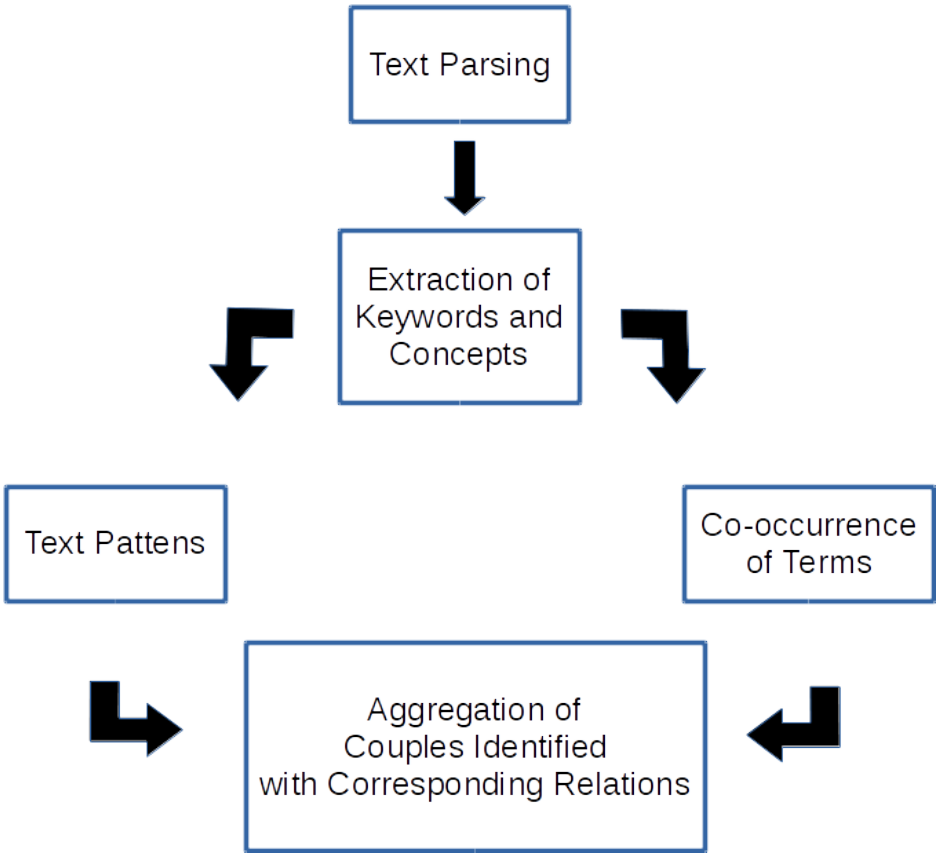



Figure 2: The specific components of the text analysis depicted in Figure 1.

- Ontology and external sources: two specific ontologies have been defined, according to the semantic properties of concepts, as well as probabilistic and statistical terms. This is discussed in Section 3.

- A text analysis component as depicted in Figure 2, which includes the identification of specific keywords (i.e., probabilistic and statistical terms used in this article), as well as general concepts, along with any other relevant semantic information, including synonyms, antonyms, meronyms, etc. This is discussed in Section 3. In particular,

  - Specific text patterns have been defined for extracting probabilistic relationships between concepts, as discussed in Section 4. This approach is also reinforced by the augmentation of co-occurrence

properties to such concepts.

  - The output of the extraction includes pairs of concepts linked by a probabilistic relationship. These can be of probabilistic *dependence* or *independence* type, as discussed in Section 4. Note that if the extraction is not successful in identifying a specific relationship between any two concepts, then it is defined *unknown*. All the above information can be suitably aggregated to enable a full investigation of their types.
  - Finally, following an appropriate data aggregation, any new concept and probabilistic information relevant to the context described in this article, is merged with the ontologies.

- The output of the text analysis is a multi-edge network where its nodes are associated with the extracted concepts and each edge represents the group of relationships between nodes. The topological properties of the networks are subsequently analysed to assess probabilistic information associated with them. As discussed in Section 5, the edges are reassessed so that they are grouped into the following categories

  - Directed edges provided that the directions of the corresponding relations have been identified. This could include a $2-$loop between two nodes, say $A$ and $B$, if a double relation is present, such as $A$ *causes* $B$ and $B$ *causes* $A$.
  - Undirected if the corresponding relation is unknown.
  - Finally, an independence relation resulting to no edge between the corresponding nodes. However, the opposite does not hold as a lack of an edge does not necessarily imply an independence relations, as discussed later on. As a consequence, even though an independence relation has no connecting edges, it will be different from the case of no edges due to no relations extracted.

- Following a consistency checking process, potentially various permissible versions of BNs are created, based on the extracted network.

- Finally, the visualisation of the BNs identified, as well as knowledge discovery, will be carried out.

Table 1: A small selection of the relation terms as in Section 3.

| **Relation Terms** |
|---|
| cause |
| activate |
| impel |
| inspire |
| excite |
| quicken |
| rouse |
| stimulate |
| influence |
| determine |
| likely |
| probable |
| independent |
| associated |
| disconnected |
| separated |
| excluded |

## 3. Ontology Definition

Although the method introduced in this work is applicable to a variety of contexts, our validation process is specifically focused on textual sources from the biomedical sector. As a consequence, two ontologies were defined to contain biomedical terms. More specifically, these were based on the hierarchical semantic properties of suitable keywords, which include:

- Probabilistic and statistical relation terms
- Biomedical concepts

The former were automatically extracted from Wikipedia based on a general query related to statistical and probabilistic types of relations. Their definitions were subsequently analysed to identify further concepts, which are semantically related. This process generated over 2300 terms, which tend to occur in sentences that indicate information related to uncertainty and probabilistic states. Table 1 shows a small selection.

More specifically, these keywords are associated with dependency and independence relations, such as *"lung cancer is associated with tobacco smoking"*, or *"an increase in aspirin assumption has been shown to be independent from cancer occurrence"*. Furthermore, (English based) WordNet (Fellbaum C, 1998) was utilised to identify keywords that were considered to be semantically equivalent. In particular, such equivalence is based on whether terms belong to the same semantic class, or *synsets* (Fellbaum C, 1998).

This was also crucial in minimising the negative effects of ambiguity based on the assumption that keywords within the same synset are deemed equivalent.
Biomedical concepts were found in the *Open Biological and Biomedical Ontologies* (The Open Biological and Biomedical Ontologies, 2014). This is a collaborative research effort that identifies a set of principles and interoperable reference ontologies in the biomedical domain. These are encoded into files by using a variety of formats including `obo` and `rdf` (DuCharme B, 2011). Their analysis populated a list of over $60,000$ concepts. As discussed in the following section, the extraction of probabilistic information is subsequently assessed and aggregated into our ontologies.

## 4. Text Analysis and Relation Extraction

The extraction of probabilistic information joining different concepts is crucial in the identification of the most appropriate fragments of BNs, which model the scenarios they represent. The concepts extracted from the Open Biological and Biomedical Ontologies are particularly effective in minimising the extraction of erroneous concepts, as well as those which are semantically equivalent. In fact, the biomedical field is particularly appropriate to this approach since most of the concepts are uniquely identifiable (Xu R, 2012), (Yu F, 2016). Furthermore, the extraction and assessment of relationships between biomedical concepts within the Big Data scenario has enormous potential and far-reaching benefits (Mars M, 2010).

On the other hand, a general approach would be prone to to a higher level of inaccuracy. Consider, for example "*high humidity in the atmosphere is linked to precipitation*", which suggests a connection between *high humidity* and *precipitation*. However, we might argue that *humidity* is the appropriate concept as *high* is just an attribute, or state. Similar cases are indeed subjective as they depend on the specific scenarios that are modelled.

In particular, specific disambiguation techniques (Manning CD, 1999) need to be implemented to enable a better and more general identification of the appropriate nodes in a BN by classifying and grouping together nodes referring to similar concepts. However, word synonymity and polysemy must be fully addressed, depending on the corresponding context (Sanchez-Graillet O, 2004), where a variety of supervised machine learning algorithms can be potentially utilised to facilitate this task. As discussed above, the investigation of suitable node attributes will also play an important role in the correct identification of the topology of the resulting networks. Furthermore, dependency and independence relations from general textual sources are likely to be defined by a larger set of linguistic expressions, which need to be captured by suitable text patterns. The next steps of our research involve addressing the above challenges by considering a variety of techniques, including more comprehensive test patterns and various statistical and machine learning approaches.

In this article, in order to ascertain the type of probabilistic linking among the different concepts, we consider a text pattern approach. This allows the identification of specific text fragments that capture probabilistic information between concepts (Feldman R, 2006). However, this approach tends to focus on specific patterns to optimise accuracy, due to the intrinsic complexity of human language.
The concepts and their mutual relations identified via text patterns, need to appear within the same text fragments, such as paragraphs, or if they are linked by specific and given relationships, such as items in a database. This was carried out by considering the following quintuples `(NP1, MOD, tense, keyword, NP2)` where:

- `NP1` and `NP2` are the *noun phrases*, i.e. phrases with a noun as the head word, which have to contain one or more biomedical concepts.
- `keyword` is one or more probabilistic terms contained in the ontology as per Section 3.
- `MOD` is the `keyword` *modality*. This identifies whether the corresponding sentence refers to either a relation or a 'non-relation' in probabilistic terms. In particular, this can be either `positive` or `negative` depending on whether it reinforces the existence of a probabilistic relationship, or negates it.

- `tense` refers to the tense of the action, whether `active` or `passive`. If it cannot be determined, then it will be defined as `unknown`.

These quintuples are specifically extracted by analysing the syntactic structure of sentences and text fragments. As discussed in Section 7, Python NLTK (Bird S, 2009) was used to create a prototype to tokenise, parse and extract the relevant syntactic information which are described by the different POS components.
A sentence such as "*a small consumption of alcohol is* **not** *linked with an increased risk of cancer*" suggests a lack of any probabilistic relationship between *small alcohol consumption* and *increased risk of cancer*. The modality *not* specifically determines such a state. We refer to this scenario as an *independence* state. Strictly speaking, such type of assertion might only suggest a lack of a probabilistic relation, rather than full independence. However, a full semantic analysis of this particular issue goes beyond the scope of this work, and as a consequence, we will assume that such type of assertion refers indeed to an independence relation.
An independence relation can also be directly specified by the `keyword` component of the aforementioned text patterns, such as "*the risk of cancer appears to be independent from a small consumption of alcohol*". Clearly, the `MOD` is positive, whereas `keyword` indicates the existence of an independence relationship.

The above process can be summed up as the following:

- `positive MOD` + `independence keyword` = independence (probabilistic) relation
- `positive MOD` + `dependence keyword` = dependence (probabilistic) relation
- `negative MOD` + `independence keyword` = dependence (probabilistic) relation
- `negative MOD` + `dependence keyword` = independence (probabilistic) relation

In this approach, we also focus on the direction of the corresponding relationships between concepts. Human interpretation would consider the verb tense and/or the type of keywords to ascertain the direction. However, such

type of analysis is demonstrated to be difficult in NLP systems (Blanco E, 2008). The approach used in this article attempts to identify the corresponding relation depending whether an active of passive form is found, as specified by the tense of the verb or keyword.
In future research, we are aiming to provide a better text analysis, which would include an enhanced independence extraction, as well as the direction of dependency relations (Trovati M, 2015 B).

## 5. Network Topology Extraction and Probabilistic Information Extraction

The quintuples `(NP1, MOD, tense, keyword, NP2)` described in Section 4, naturally define a network, where the concepts in `NP1` and `NP2` are connected by an edge whose type depends on the associated relations. More specifically, the network is defined as $G = G(V, E)$, where $V = \{v_i\}_{i=1}^{n}$ is the *node set* and $E = \{e_{v_i,v_j}\}_{v_i,v_j \in V}$ is the *edge set*. Usually, such networks are multi-graphs, where nodes may have more than one edge between them, with the exception of self-loops, or in other words edges starting and ending from the same node.
In fact, more than one relation between two concepts might be present, depending on the nature and size of the text corpora analysed. Therefore, each element of $E$ is itself a non-trivial set containing the relationships between pairs of nodes, which have been extracted via the text analysis.
These relations can be of *dependency*, *independency*, or *unknown* type and will be denoted as $S_D(v_i, v_j), S_I(v_i, v_j)$ and $U(v_i, v_j)$, respectively, for $v_i, v_j \in V$. In particular, edges in $S_D(v_i, v_j)$ are directed, where the direction is specified by the `tense` component in the text, or by information contained in the ontology. Note that, independence relations are not associated with a direction, even though the tense of the action may suggest otherwise.
In the next section, we will discuss the assessment of the conditional probabilities based on the topological properties of the network $G = G(V, E)$.

### *5.1. Probabilistic Information Extraction*

The estimation of the probability of a concept extracted from text corpora typically depends on the frequency of its occurrences, as well as on the topology of the network to which such concept belongs (Trovati M, 2015 A). In fact, the probability of choosing, or rather *observing* a concept $A$ is

proportional to the number of its connections. Intuitively, the higher the connectivity of one node is, the more it is assumed to be observable. This is evaluated by the ratio

$$p_T(A) = \frac{d(A)}{|E|}, \tag{1}$$

where $d(A)$ is the degree of the node $A$, and $|E|$ is the total number of edges. We then define the probability of $A$, as

$$P(A) = p_O(A)p_T(A), \tag{2}$$

where

$$p_O(A) = \frac{O_A}{O_V}, \tag{3}$$

and $O_A$, $O_V$ are the number of occurrences of concept $A$ and all the concepts in $V$, respectively. Recall that a scale-free network has a degree distribution, which asymptotically follows a power law (Albert R, 2002). In other words, the fraction $P(k)$ of nodes in the network having $k$ neighbouring nodes for large $k$, is described as

$$P(k) \approx k^{-\gamma}, \tag{4}$$

where $\gamma$ is typically between 2 and 3. Following Equations 1 and 4, if the network $G$ exhibit a scale-free structure, then we can assume that $p_T(A) = 1 - d(A)^{-\gamma}$. So Equation 2 can be written as

$$P(A) \approx \frac{O_A}{O_V}(1 - d(A)^{-\gamma}) \tag{5}$$

*5.2. Conditional Probability*

As discussed above, for two nodes $A$ and $B$, in general more than one relation might have been extracted, which could consist of dependency, independent and unknown relations, that is $S_D(A, B)$, $S_I(A, B)$ and $U(A, B)$, respectively. In particular, we make the assumption that the overall relation set between these two (connected) nodes $e_{A,B} = S_D(A, B) \cup S_I(A, B) \cup U(A, B)$ is non-empty. Furthermore, a high number of unknown relations is assumed to suggest a weaker dependency, since by definition, they are not necessarily associated with a specific relation type.

We propose the following model to assess the conditional probability

$$\begin{aligned} P(A|B) &= \Lambda\, e^{-\frac{|S_I(A,B)|}{|S_D(A,B)|+|U(A,B)|}} \\ &+ P(A)\left(1 - e^{-\frac{|S_I(A,B)|}{|S_D(A,B)|+|U(A,B)|}}\right), \end{aligned} \tag{6}$$

where

$$\Lambda = \begin{cases} 1 - Ke^{-\frac{|U(A,B)|^2+|S_D(A,B)|}{|U(A,B)|}} & \text{if} \quad |U(A,B)| \neq 0 \\ 1 & \text{otherwise} \end{cases}$$

for a constant $K > 0$.
In particular, Table 2 depicts all the possible outcomes of the extraction and how to subsequently interpret and determine the conditional probability $P(A|B)$. The joint probability is therefore

$$\begin{aligned} P(A,B) = P(B)P(A|B) &= \Lambda\, P(B)\, e^{-\frac{|S_I(A,B)|}{|S_D(A,B)|+|U(A,B)|}} \\ &+ P(B)P(A)\left(1 - e^{-\frac{|S_I(A,B)|}{|S_D(A,B)|+|U(A,B)|}}\right) \end{aligned} \tag{7}$$

Note that the factor $\Lambda$ measures the strength of dependency relations compared to the number of unknown relations. If $|S_D(A,B)| << |U(A,B)|$, the above will tend to $\left(1 - Ke^{-|U(A,B)|}\right)$. This is the value that we assume to be associated to a relation characterised by unknown relationships. Depending on the number of these unknown relationships, its range is between $[1 - K/e, 1)$, for $|U(A,B)| = 1$, and $|U(A,B)| \to \infty$, respectively. The value of the constant specifies the lower bound of the range, which is associated with the conditional probability given by one and only one relation between $A$ and $B$ of unknown type. In this article, we assume that such lower bound is 0.1, which yields $K = 0.9e \approx 2.45$. This assumption was agreed upon with a group of experts, who have contributed to the validation process as discussed in Section 7.

*5.2.1. Properties of Conditional Probability*

A well know property of conditional probability is the chain rule

$$P(\bigcap_{k=1}^{n} A_k) = \prod_{k=1}^{n} P(A_k|\bigcap_{j=1}^{k-1} A_j), \tag{8}$$

Table 2: The values of the conditional probabilities and relation types for the different values of $|S_I(A,B)|, |S_D(A,B)|$, and $|U(A,B)|$.

| $P(A\|B)$ | $\|S_I(A,B)\|$ | $\|S_D(A,B)\|$ | $\|U(A,B)\|$ | **Relation Type** |
|---|---|---|---|---|
| Evaluated with (6) | $= 0$ | $= 0$ | $\neq 0$ | Dependency |
| Evaluated with (6) | $= 0$ | $\neq 0$ | $\neq 0$ | Dependency |
| $P(A)$ | $\neq 0$ | $= 0$ | $= 0$ | Independence |
| $P(A)$ | $\neq 0$ | $= 0$ | $\neq 0$ | Independence |
| Evaluated with (6) if dependency. $P(A)$ if independent | $\neq 0$ | $\neq 0$ | $= 0$ | If $\|S_D(A,B)\| > \|S_I(A,B)\|$: Dependency If $\|S_D(A,B)\| < \|S_I(A,B)\|$: Independence |
| Evaluated with (6) if dependency. $P(A)$ if independent | $\neq 0$ | $\neq 0$ | $\neq 0$ | If $\|S_D(A,B)\| > \|S_I(A,B)\|$: Dependency If $\|S_D(A,B)\| < \|S_I(A,B)\|$: Independence |
| Evaluated with (6) | $= 0$ | $\neq 0$ | $\neq 0$ | Dependency |

which is used to define the joint probability of more than on concept. Equation 8 can be written explicitly as

$$P(A_1, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2A_1)\cdots P(A_n|A_{n-1}\cdots A_1). \quad (9)$$

In this article, we define $P(A_n|A_{n-1}\cdots A_1)$ similar to Equation 6. However, since we have to consider the influence of the nodes $A_{n-1}, \cdots, A_1$ on $A_n$, we will only consider the minimum of each of their contribution in terms of dependency, independence and unknown relations. Namely,

$$P(A_n|A_n\cdots A_1) =$$

$$\left(1 - Ke^{-\frac{|U(A_n,\cdots,A_1)|^2+|S_D(A_n,\cdots,A_1)|}{|U(A_n,\cdots,A_1)|}}\right) e^{-\frac{|S_I(A_n,\cdots,A_1)|}{|S_D(A_n,\cdots,A_1)|+|U(A_n,\cdots,A_1)|}} + P(A_n)\left(1 - e^{-\frac{|S_I(A_n,\cdots,A_1)|}{|S_D(A_n,\cdots,A_1)|+|U(A_n,\cdots,A_1)|}}\right), \quad (10)$$

where

$$\begin{aligned} S_D(A_n,\cdots,A_1) &= \min\{S_D(A_n,A_{n-1}),\ldots,S_D(A_n,A_1)\} \\ S_I(A_n,\cdots,A_1) &= \min\{S_I(A_n,A_{n-1}),\ldots,S_I(A_n,A_1)\} \\ U(A_n,\cdots,A_1) &= \min\{U(A_n,A_{n-1}),\ldots,U(A_n,A_1)\}. \end{aligned}$$

Furthermore, analysing Equation 6, we can see that

- If $A$ and $B$ are independent, then $S_I(A,B)$ is large with respect to $S_D(A,B)$ and $U(A,B)$, which implies that $e^{-\frac{|S_I(A,B)|}{|S_D(A,B)|+|U(A,B)|}} \to 0$, and subsequently $P(A|B) \approx P(A)$.
- The assertion $A \supset B$ is assumed to signify that $B$ is connected to $A$ with a large number of $S_D(A,B)$, both globally and relatively to $S_I(A,B)$ and $U(A,B)$. This implies that $e^{-\frac{|S_I(A,B)|}{|S_D(A,B)|+|U(A,B)|}} \to 1$, and so $P(A|B) \approx 1$.
- For two nodes $A_1$ and $A_2$ such that $A_1 \cap A_2 = \emptyset$, then we assume that these two nodes have no mutual connection. This implies that $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B)$.
- And finally, $P(A|B) \neq P(B|A)$, which is obvious from Equation 6.

All the above is consistent with the general properties of conditional probability.

### *5.3. Identification of Consistent BNs*

As discussed above, a BN is a directed, acyclic graph where edges are associated with a dependence relations usually characterised by a conditional probability. Furthermore, these depend on the direction of the corresponding edge as, for example, $P(A|B)$ implies a directed edge $B \to A$, and vice-versa. Since the conditional probability values are extracted from text, and following the method discussed above, edges can exhibit the following types:

- A *single* direction, such as $A \leftarrow B$, or $A \rightarrow B$. This can be either extracted from the text, or determined from the ontology.

- A *double* direction, in the sense that both $P(A|B)$ and $P(B|A)$ have been extracted. This is only possible if the directions have been identified from text. Different texts, for example, could describe a variety of information, such as "*heart disease will influence lifestyle changes*" and "*lifestyle changes can contribute to an increased likelihood of heart disease*". Clearly, these two statements are not contradictory and fully explainable within their contexts. However, a text mining analysis might interpret them as a double dependency between *lifestyle changes* and *heart disease*. As a consequence, in determining which direction the linking edge should have, it might not be directly possible to ascertain whether they refer to different contexts, as well as whether the extraction was indeed accurate. In this article, we therefore assume both directions have to be considered. As discussed above, this is opposite to knowledge from the ontology, which is assumed 'certain' with a well defined dependency direction.

- A *strong independence*, is identified by textual analysis as an independent relation, or from the ontology, should this be specified. This is to diversify the lack of an edge between two concept, which does not necessarily imply independence between them, as this might be due to error in the the extraction, or noise in the data (Yang X, 2014).

- Finally, undirected edges correspond to unknown relations, which have not been successfully classified. This can only be determined from textual sources.

The ontology described in Section 3 provides the information, which has priority over any knowledge extracted from text. As a consequence, if any extracted relationship has a different connotation in the ontology, the former will be discarded and the latter will be assumed to be correct.

### *5.4. Consistency Checking and Creation of Copies of a BN*

Since BNs are directed acyclic graphs, only directions which define consistent BNs need to be considered.
There is a wealth of literature on loop removal within BNs to provide consistency, see (Ben-Gal I, 2007) for an overview. Therefore, in order to ensure

consistency, networks with no cycles will be kept. These are defined as the set of *copies* $\mathcal{G} = \{G_i\}_{i=1}^{k}$ of $G$, where $G_i = G_i(V, E_i)$ is defined by the method above. Note that in general, $|E| \geq |E_i|$ for $i = 1, \ldots, k$.
In this article, the approach to identify copies is based on the following assumptions:

**Unknown edges:** each direction should be investigated. If neither of them provides a consistent network, then such edge will be removed.

**Double directed edges:** each direction should be investigated separately, and only directions that are consistent will be kept. If neither directions provide consistency, then the corresponding copy will be discarded. Recall that this type of edges are associated with relations extracted from textual sources.

**Strong independent edges:** no actions are possible, as we cannot add an edge between the corresponding nodes. This will come from both the extraction from text and the ontology.

**Directed edges determined from the ontology:** as discussed above, these types of edges are deemed certain and therefore they cannot be removed.

Algorithm 1 describes the creation of the set of copies $\mathcal{G}$.

In general, some edges might have values of their extracted conditional probability too small, and their existence should be therefore assessed. Usually, such assessment is carried out manually depending on the context. In this article, this process focuses on the idea that the threshold depends on the distribution of the values of the conditional probability and how evenly they are distributed over the edges. Therefore, let

$$\bar{P}_G = \frac{w(A_i, A_j)P(A_i|A_j)}{\sum_{e_{A_i,A_j} \in E} w(A_i, A_j)}, \tag{11}$$

where $A_i, A_j \in V$, and $w(A_i, A_j)$ refers to the number of edges with the same conditional probability value.
Therefore, only edges whose conditional probability is greater or equal to $\bar{P}_G$ will be kept.
In general, for a network $G = G(V, E)$, where $E$ includes $d$ double directed edges and $u$ unknown ones, it is straightforward to see that there are at most

**Algorithm 1** Creation of Copies of BNs

Let $G = G(V, E)$ be the network defined above.
Define $E_D$ and $E_U$ as the subsets of $E$ containing the dependency and unknown relations, respectively.
Let $e_{v_i,v_j} \in E$ be an edge between the nodes $v_i$ and $v_j \in V$.
**for** $e_{v_i,v_j} \in E_D$ consider both directions **do**
  **for** $e_{v_i,v_j} \in E_U$ consider the three possible directions **do**
    **if** the corresponding sub-network $G_i$ is consistent **then**
      Add $G_i$ to $\mathcal{G}$
    **else**
      Discard $G_i$
    **end if**
  **end for**
**end for**
**return** $\mathcal{G}$

$2^d 3^u$ copies of corresponding BNs. Clearly, this is not computationally efficient, especially when considering large networks. However, in the analysis carried out in this article, the number of double directed edges and the unknown relation is smaller than those identified via the ontology, due to the context of our analysis.

## 6. Knowledge Discovery

As discussed above, a lack of an edge between two nodes does not imply an independence relation, as no explicit relation might not have been identified from the data. Unless two disconnected nodes have been explicitly shown to be independent, investigating the existence of a relation between them by adding an edge is, in principle, a permissible action. However, such action should not invalidate the consistency of the corresponding copy.

The factors that influence the feasibility of a direct connection include (Trovati M, 2015 A):

- The number of paths between $A$ and $B$,
- The type of relations of the corresponding edges, that is dependency and unknown,

- The strength of such hypothetical connection in terms of the corresponding conditional probabilities.

Let $\mathcal{P}^l_{A,B}$ be the set of all the paths with length $l \geq 2$ and $\mathcal{P}_{A,B} = \bigcup_l \mathcal{P}^l_{A,B}$ be the set of all paths between $A$ and $B$.
Let

$$R^l(A,B) = \frac{1}{\log(l)} \max_{p_{A,B} \in \mathcal{P}^l_{A,B}} \left\{ \frac{1}{l} \sum_{i=1}^{l} P(x_i|x_{i-1}) \right\}, \tag{12}$$

where $x_i$ and $x_{i-1}$ are two consecutive nodes along a path $p_{A,B} \in \mathcal{P}^l_{A,B}$. We then define the *relational strength* $R(A,B)$ as

$$R(A,B) = \max_l \{R^l(A,B)\}. \tag{13}$$

Note that $R(A,B) \neq R(B,A)$, and neither of them is defined if $\mathcal{P}_{A,B} = \emptyset$, or $\mathcal{P}_{B,A} = \emptyset$, respectively.
Loosely speaking, $R(A,B)$ measures the (semi-directed) "closeness" between $A$ and $B$, and it is not necessarily linked with $P(A|B)$. Furthermore, shorter paths between $A$ and $B$ will imply a higher $R(A,B)$ value.
Typically, $R(A,B)$ may depend on the copy of the BN, which is considered. If the existence of a link between two nodes is supported by different copies, then we can assume this suggests a strong closeness between the two nodes. In other words, if $R(A,B)$ is within an interval for several copies, then the corresponding nodes are likely to be connected. However, not all copies may be an accurate representation of a system, and this largely depends on the modeller's interpretation. Therefore, in evaluating Equation 13 we should distinguish between choosing specific copies, as opposed to considering all of them.

## 7. Implementation and Evaluation Results

The evaluation utilised in this work is based on an annotated dataset, which was created via a variety of tools, including (but not limited to) Python NLTK (Bird S, 2009). The analysis was carried out using a single Linux machine, with an I7 processor and 16 Gb of memory. Although a discussion on the optimisation of the computational time goes beyond the scope of this work, the extraction and analysis was completed in approximately 10 hours for a non-parallelised approach, as opposed to just over 7 hours based on

parallel processing provided by the `multiprocessing` Python library. Subsequently, this was manually assessed and annotated by a group of experts in the medical sector to provide a gold-standard dataset.

Part of the above dataset was based on

- 800 articles and 1000 abstracts were randomly selected and downloaded from PubMed Text Mining tools in `xml` format, based on specific keywords as discussed in Section 7.1.

- These were POS tagged, and analysed via text patterns as described in Section 4. The overall POS and concept extraction was manually assessed by considering a sample of approximately 120 abstracts and 40 articles, which showed 89% recall and 79% precision. As discussed earlier, such good results are also influenced by the low level of ambiguity, which is typical of biomedical literature.

- The output of the above analysis was a list of triples

  ```
  [concept_1, concept_2, Rel]
  ```

  where

  - `concept_1` and `concept_2` refer to couples of biomedical concepts, and
  - `Rel` is either 1 if we have a probabilistic dependence relation, or $-1$ if it refers to an independence one. If `Rel` is of unknown type, then it will be set to 0.

The network extracted was analysed using the approach introduced in (Trovati M, 2015 A), which confirmed it has a scale-free structure. Finally, the visualisation of the BN copies was carried out in Matlab.

The evaluation focused on the following components

- First the conditional probabilities were assessed by a group of experts in the medical field and benchmarked with a variety of methods, as discussed in Section 7.1.

- Secondly, the extraction of suitable copies of BNs was evaluated. This was carried out by comparing our results with a manual creation of BNs based on the experts' findings, as described in Section 7.2.

*7.1. Evaluation of the Conditional Probability*

We evaluated our approach by considering approximately 800 articles and 1000 abstracts from the biomedical sector.
First of all, we considered the method introduced in (Theobald M, 2009), where conditional probabilities between drugs, diseases, and genes are identified from abstracts in PubMed to construct $n-$way Bayesian networks based on co-occurrence statistics. This, however, might affect the precision of the analysis as co-occurrence by itself is in general unable to distinguish between legitimate statistical relationships. Furthermore, their proposed method is based on pharmGKB, and no direct relations can be discovered unless specifically mentioned. In this part of the evaluation, the abstracts and articles were selected based on keywords related to (Theobald M, 2009) including the concepts shown in Table 3. More specifically, the gold-standard dataset discussed above was utilised, which consists of a processed and annotated textual dataset. Note that the values highlighted in bold-face are those that fall closest to or within the interval identified by the manual assessment carried out by experts in the medical sector.

The concepts extracted and their mutual relationships defined a network with over 740 nodes and 850 edges. Using the method discussed in (Trovati M, 2015 A), we ascertained that it exhibits a scale-free structure, with $\gamma \approx 2.1$. Figure 3 depicts the distribution of node connectivity.

A rigorous comparison with (Theobald M, 2009) would require the same dataset, which is not the case in this work, and as such, the results do not imply that either method is more accurate. However, the aim of this part of the evaluation is to demonstrate that the model discussed in Section 5.2 and in particular Equation 10, produce meaningful results.

We also considered the well-known pharmacogenic associations described in PharmGKB (PharmGKB, 2016), which consists of approximately 664 associated couples, where a small selection of them is reported in Table 4. Given these relationships have been investigated and fully established, we assumed that the joint probability of any of the couples should be above 0.8. We ran our method over the corpus discussed above, with the exception that we expanded the number of abstracts to approximately 1250 specifically selected to include a variety of pharmacogenic associations. We obtained that 69% of them were identified with a joint probability above the threshold.

In (Jurca G, 2016), a method for analysing relationships between different typologies of cancer and specific genes is introduced. Using a similar

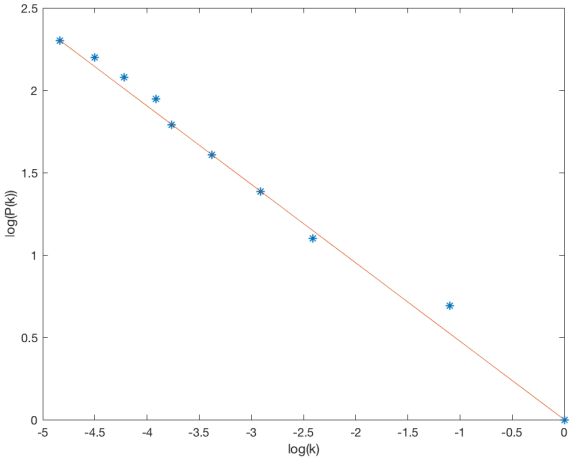



Figure 3: The graph of a selection of values of the ratio of nodes with degree $k$. The line is associated with $\gamma = 2.1$, showing it accurately describes the scale-free structure of the extracted network.

evaluation, a query based on "breast cancer" was carried out to identify articles with an abstract, title, authors, and a journal name during October 2014. Furthermore, these articles were grouped according to the geographical location by extracting the main author's affiliation.
Table 11 in (Jurca G, 2016) shows the top 10 genes mentioned by articles grouped by country, and in our evaluation we considered United States, United Kingdom, Germany and France, as depicted in Table 5. As above, we set a 0.8 joint probability of "breast cancer" and any of the genes for each country. Our method correctly identified all the gene names for United States and France, whereas PGR and KRT75, BRCA1 for United Kingdom and Germany, respectively were not successfully captured.

In (Jochim C, 2016), a method to extract probabilities and risk events from biomedical texts is discussed. In particular, the conditional probability of two events $A$ and $B$ is assessed as

$$P(A|B) = \frac{\text{PubMed}(A \cap B)}{\text{PubMed}(B)}, \tag{14}$$

where PubMed$(A \cap B)$ and PubMed$(B)$ are the number of hits from PubMed queries related to both $A$ and $B$ and $B$, respectively. Similarly to (Jochim C, 2016), we randomly selected 5000 abstracts identified by a "breast cancer"

Table 3: The results of our method and those produced by (Theobald M, 2009). The highlighted values in bold-face are those closest to, or within the range specified by the experts.

| **Type** | **Our Approach** | **Approach in (Theobald M, 2009)** | **Experts' Assessment** |
|---|---|---|---|
| $p$(antidepressants\| affective disorders, GNB3) | **0.41** | 0.33 | $0.45 - 0.55$ |
| $p$(mercaptopurine\| azathioprine, thioguanine, TPMT) | **0.72** | 0.84 | $0.7 - 0.8$ |
| $p$(thioguanine\| azathioprine, mercaptopurine TPMT) | 0.65 | **0.73** | $0.7 - 0.8$ |
| $p$(azathioprine\| mercaptopurine, thioguanine, TPMT) | **0.77** | 0.89 | $0.6 - 0.75$ |
| $p$(azathioprine\| thioguanine, TPMT) | **0.58** | 0.86 | $0.65 - 0.75$ |
| $p$(thioguanine\| azathioprine, TPMT) | **0.65** | 0.44 | $0.5 - 0.6$ |
| $p$(salmeterol\| Asthma, ADRB2) | **0.22** | 0.07 | $0.1 - 0.2$ |
| $p$(salbutamol\| Asthma, ADRB2) | **0.28** | 0.16 | $0.2 - 0.3$ |

query. Furthermore, specific concepts related to breast cancer were identified via the Metathesaurus provided by the Unified Medical Language System (UMLS) (UMLS, 2017). These include

Table 4: A small selection of well-known pharmacogenic associations as in (PharmGKB, 2016)

| **Drug** | **Gene** |
|---|---|
| Abacavir | HLA-B |
| Abiraterone | CYP17A1 |
| Abiraterone | CYP1A2 |
| Acenocoumarol | CYP2C9 |
| Acenocoumarol | VKORC1 |
| Acetaminophen | CYP1A2 |
| Acetaminophen | CYP2D6 |
| Afatinib | ABCB1 |
| Afatinib | ABCG2 |
| Afatinib | CYP1A2 |

- Malignant neoplasm of breast,
- Breast Carcinoma,
- Noninfiltrating Intraductal Carcinoma,
- Mammary Ductal Carcinoma,
- Malignant neoplasm of female breast,
- Carcinoma of Male Breast,
- Carcinoma breast stage I,
- Breast cancer stage II,
- Breast cancer recurrent,
- Lobular carcinoma.

We subsequently assumed that the conditional probability of breast cancer given any of the above concepts is higher than 0.7, as in (Jochim C, 2016). Our approach correctly identified the conditional probability except for lobular carcinoma, as $P($ breast cancer$|$ lobular carcinoma$) \approx 0.45$. Although our method and experimental setting are slightly different from (Jochim

Table 5: The gene names associated with "breast cancer" grouped by country as in (Jurca G, 2016).

| **Country** | **Gene Name** | **Country** | **Gene Name** |
|---|---|---|---|
| United States | ESR1 | United Kingdom | ESR1 |
| | ERBB2 | | ERBB2 |
| | EGF | | CYP19A1 |
| | PGR | | EGF |
| | BRCA1 | | BRCA1 |
| | CDKN2A | | CDKN2A |
| | SLC20A2 | | PGR |
| | TKT | | BRCA2 |
| | ACAD9 | | SLC20A2 |
| | CYP19A1 | | INS |
| Germany | ERBB2 | France | ESR1 |
| | ESR1 | | ERBB2 |
| | PGR | | PGR |
| | EGF | | CDKN2A |
| | CDKN2A | | BRCA1 |
| | SLC20A2 | | EGF |
| | BRCA1 | | SLC20A2 |
| | CYP19A1 | | TKT |
| | KRT75 | | CYP19A1 |
| | | | CTSD |

C, 2016), the evaluation demonstrated a comparable, if not enhanced conditional probability extraction.

In the next validation component, the following concepts were selected:

- *cancer* and *smoking*
- *heart disease* and *beta-blocker*
- *organophosphate* and *pyrethroid*
- *symptom* and *disease*

These were used to assess the methods described in Section 5.2, as described in Table 6 and compared with a probability range given by the group of medical experts.

Table 6: Evaluation of the conditional probabilities.

| **Concept** $A$ | **Concept** $B$ | $P(A\|B)$ | **Expert's Assessment** |
|---|---|---|---|
| Cancer | Smoking | 0.75 | $0.8 - 0.9$ |
| Heart Disease | Beta-Blockers | 0.5 | $0.4 - 0.6$ |
| Organophosphate | Pyrethroid | 0.21 | $0.2 - 0.5$ |
| Symptom | Disease | 0.89 | $0.8 - 0.9$ |

*7.2. Evaluation of the Extraction of Copies of BNs*

The evaluation of the extraction of relevant BNs followed a similar approach. In fact, we first queried PubMed based on the following keywords:

- Liver disorder
- Heart disease
- Mental disorder
- Breast cancer

For each keyword, we further selected approximately 200 abstracts from PubMed, which were subsequently analysed. The networks generated each had an average of approximately 175 nodes and 350 edges. However, we only considered a small proportion for each of them as, for part of the validation process, they had to be compared with a manual extraction. Therefore, we only considered selection defined by a small sub-group of nodes, as described in Table 7.

Examples of dependency and independence relations, which were extracted include the following:

- Independence relations
  - *"State anxiety describes the person's feelings at a particular time and under particular conditions, whereas trait anxiety is independent of conditions and reflects stable personality characteristics and generalised feelings"*
  - *"Our theory states that anxiety is independent of intelligence"*
  - *"In addition, the present results show that sex- and age-related changes in immunoglobulin concentrations are independent of potential confounders such as smoking, alcohol consumption and common metabolic abnormalities".*

Table 7: The concepts analysed as discussed in Section 7.2

| **Liver Disorder** | **Heart disease** |
|---|---|
| Alcohol Abuse | Diet |
| Liver disorder | Physical Exercise |
| Body Hair Loss | Obesity |
| Alkaline Phosphatase | Triglycerides |
| Total Bilirubin | Cholesterol |
| ESR | Smoking |
| LE cells | High Blood Pressure |
| Total Proteins | Atherosclerosis |
| Direct Bilirubin | Heart Disease |
| JaundiceItching | ECG |
| Alpha Globulin | Angina |
| Beta Globulin | Myocardial Infarction |
| Gamma Globulin | |
| IgA | |
| IgB | |
| Smooth Muscle Antibodies | |
| Anti-mitochondrial Antibodies | |
| Musculo-Skeletal Pain | |

| **Mental disorder** | **Breast cancer** |
|---|---|
| Mental Disorder | Breast Cancer |
| Anxiety | Age |
| Sleep Disorder | Pain |
| Mood Disorder | Menarche Age |
| Psychotic Disorder | Nipple Discharge |
| Personality Disorder | Asymmetry |
| Dissociative Disorder | Calcification |
| Schizophrenia | Calcification Density |
| Delusional Disorder | Mass |
| Catatonic Schizophrenia | |
| Paranoid Schizophrenia | |

- Dependency relations
  - *"Diet plays an important role in the cause and the prevention of*

Table 8: Properties of the network extracted in the validation process.

| **Network by Keyword** | **Number of Possible Networks** | **Number of Copies** | **Values of $\bar{P}_G$** |
|---|---|---|---|
| Liver Disorder | 141 | 27 | 0.55 |
| Heart disease | 113 | 19 | 0.66 |
| Mental disorder | 101 | 13 | 0.49 |
| Breast cancer | 37 | 11 | 0.33 |

*heart disease"*

- *"The literature indicates that anxiety comorbidities are prevalent in schizophrenia and conventional treatment for anxiety can help alleviate the symptoms in those patients"*
- *"Research suggests that persons with schizophrenia tend to experience significant levels of anxiety".*

Table 8 shows the number of possible networks found for each network, the number of copies, and the corresponding values of $\bar{P}_G$. The latter was used to remove edges below its corresponding value.

All the copies were subsequently analysed by the experts to assess the proportion of BNs, which adequately model the scenarios corresponding to the different keywords. Although these copies were not ranked separately, the experts confirmed that an average of 65% of them were considered as valid. In particular, Figures 4 – 7 depict the full networks extracted and the copy which was unanimously considered as the optimal one by the experts who carried out the manual validation, confirming that at least one copy was deemed correct.

In the final part of of the validation, for each of the above networks we removed one of the existing edges referring to well documented dependencies to assess whether the method described in Section 6 correctly identified their existence. These are as follows

- Liver Disorder: "Total Bilirubin → Jaundice",
- Heart Disease: "Physical Exercise → Obesity",

Table 9: The comparison of the number of copies with manual extraction.

| **Network by Keyword** | **Removed Edge** | **Edge Correctly Identified?** |
|---|---|---|
| Liver Disorder | Total Bilirubin → Jaundice | No |
| Heart disease | Physical Exercise → Obesity | Yes |
| Mental disorder | Personality Disorder → Psychotic Disorder | Yes |
| Breast cancer | Breast Cancer → Calcification | Yes |

- Liver Disorder: "Personality Disorder → Psychotic Disorder",
- Liver Disorder: "Breast Cancer → Calcification".

As discussed in Section 6, the relational strength as defined in Equation 13, depends on a variety of parameters, including the modeller's preferences and the selected copies. In this work, we have set the threshold value above which a relation is indeed discoverable and present, as 0.3. This was discussed and agreed upon with the group of experts, who took part in the evaluation. Table 9 shows the evaluation results, and three out of four edges were identified correctly.

## 8. Conclusions and Future Directions

In this work, we have introduced a novel method for extracting, assessing, and evaluating fragments of BNs from textual sources, based on grammar and lexical properties, as well as on the topological properties of the network extracted. Furthermore, the ability to assess knowledge which is not directly specified, allows the discovery of relations between concepts in the corresponding BNs. The evaluation clearly shows the accuracy and potential of the proposed approach. Our knowledge and representation of the world relies heavily on unknown parameters, which are based on *a priori* knowledge and inadequate certainty of the particular scenario to be modelled. Furthermore, the accuracy and availability of the underlying information is also a crucial aspect to be addressed. BNs provide a very powerful tool for a broad

range of applications, with particular emphasis on cause and effect modelling in a wide variety of domains. This clearly has a number of far reaching applications, which have important implications in multiple contexts and research fields. In particular, we are aiming to embed this effort into a wider line of research, where text corpora from different sources and contexts are fully analysed, and subsequently integrated with a deeper semantic analysis to allow a full network investigation and a more comprehensive knowledge discovery process.

## Bibliography


Trovati M and Bessis N An influence Assessment Method Based on Co-Occurrence for Topologically Reduced Big Data Sets. *Soft Computing*, Springer Berlin Heidelberg, 2015

Ji J, Hu R, Zhang H and Liu C A Hybrid Method for Learning Bayesian Networks Based on Ant Colony Optimization. *Applied Soft Computing*, Volume 11, Issue 4, June 2011, Pages 3373-3384

Cruz-Ramirez N, Acosta-Mesa H G, Carrillo-Calvet H, and Barrientos-Martinez R E Discovering Interobserver Variability in the Cytodiagnosis of Breast Cancer Using Decision Trees and Bayesian Networks. *Applied Soft Computing*, Volume 9, Issue 4, September 2009, Pages 1331-1342

Dojer N, Bednarz P, Podsiadlo A, and Wilczynski B BNFinder2: Faster Bayesian Network Learning and Bayesian Classification. *Bioinformatics Applications Note* Vol. 29 no. 16, pages 2068–2070, 2013

Watts D J and Strogatz H S Collective Dynamics of Small-World Networks. *Nature*, 393, pp. 440-442, 1998

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

Dias G, Mukelov R, Cleuziou G Mapping General-Specific Noun Relationships to WordNet Hypernym/Hyponym Relations. *Proceedings of the 16th International Conference*, EKAW 2008, Acitrezza, Italy, pp 198–212, 2008

Zipf G K Human Behavior and the Principle of Least Effort. *Addison-Wesley Press*, 1949

Fellbaum C WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press. *MA: MIT Press*, 1998

The Open Biological and Biomedical Ontologies. Available from http://www.obofoundry.org/. [15 March 2014]

DuCharme B Learning SPARQL. *O'Reilly Media*, 2011

Mars M and Scott R E Global E-Health Policy: A Work In Progress. *Health Affairs*, vol. 29 no. 2 237-243, 2010

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

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

AnHai D, Raghu R and Shivakumar V Managing Information Extraction: State of the Art and Research Directions. *SIGMOD '06: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, pages 799–800, 2006

Yu F, Moh M and Moh TS Towards Extracting Drug-Effect Relation from Twitter: A Supervised Learning Approach. *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity)*, 2016

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

Xu R and Wang Q A Knowledge-Driven Conditional Approach to Extract Pharmacogenomics Specific Drug-Gene Relationships from Free Text. *Journal of Biomedical Informatics*, vol. 45, pages 827–834, 2012

Rance B, Doughty E, Demner-Fushman D, Kann M G and Bodenreider O A Mutation-Centric Approach to Identifying Pharmacogenomic Relations in Text. *Journal of Biomedical Informatics*, vol. 45, pages 835–841, 2012

Theobald M, Shah N and Shrager J Extraction of Conditional Probabilities of the Relationships Between Drugs, Diseases, and Genes from PubMed Guided by Relationships in PharmGKB. *Summit on Translational Bioinformatics*, pages 124–128, 2009

Pubmed Available from http://www.ncbi.nlm.nih.gov/pubmed. [15 June 2016]

PharmGKB Available from https://www.pharmgkb.org/. [15 June 2016]

Raghuram S, Xia Y, Palakal M, Jones J, Pecenka D, Tinsley E, Bandos J and Geesaman J Bridging Text Mining and Bayesian Networks. *IEEE Computer Society*, pages 298–303, 2009

Blanco E, Castell N and Moldovan D I Acquiring Bayesian Networks from Text. *LREC,* 2008

Trovati M, Castiglione A, Bessis N and Hill R A Kuramoto Model Based Approach to Extract and Assess Influence Relations. *Proceedings of the International Symposium on Intelligence Computation and Applications*, 2015

Sanchez-Graillet O and Poesio M Acquiring Bayesian Networks from Text. *LREC*, 2004

Manning C D and Schütze H Foundations of Statistical Natural Language Processing. *The MIT Press*, 1999

MeSH Available from https://www.nlm.nih.gov/mesh/meshhome.html [15 June 2016]

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

Ben-Gal I Bayesian Networks. Chapter in *Ruggeri F, Faltin F and Kenett R, Encyclopedia of Statistics in Quality and Reliability*, Wiley & Sons, 2007

Yang X and Mao K Multi Level Causal Relation Identification Using Extended Features. *Expert Systems with Applications*, Volume 41, Issue 16, Pages 7171-7181, 2014

Bird S, Loper E, and Klein E Natural Language Processing with Python. *O'Reilly Media Inc.*, 2009

Jensen F V Bayesian Networks. *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol 1 Issue 3, pages 307 –315, 2009

Kuipers B J and Kassirer J P Causal Reasoning in Medicine: Analysis of a Protocol. *Cognitive Science*, Vol. 8, pages 363–385, 1984

Kuipers B J Commonness Reasoning About Causality: Deriving Behaviour from Structure. *Artificial Intelligence*, Vol. 24, pages 169–203, 1985

Pearl J Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. *Morgan Kaufmann Publishers, Inc.*, 1998

Jurca G, Addam O, Aksac A, Gao S, Ozyer T, Demetrick D and Alhajj R Integrating Text Mining, Data Mining, and Network Analysis for Identifying Genetic Breast Cancer Trends. *BMC Res Notes*, 2016

Jochim C, and Sacaleanu B and Deleris L Risk Event and Probability Extraction for Modelling Medical Risks. *Proceedings of AAAI Fall Symposium*, 2014

Unified Medical Language System: UMLS Available from https://uts.nlm.nih.gov/home.html [01 March 2017]

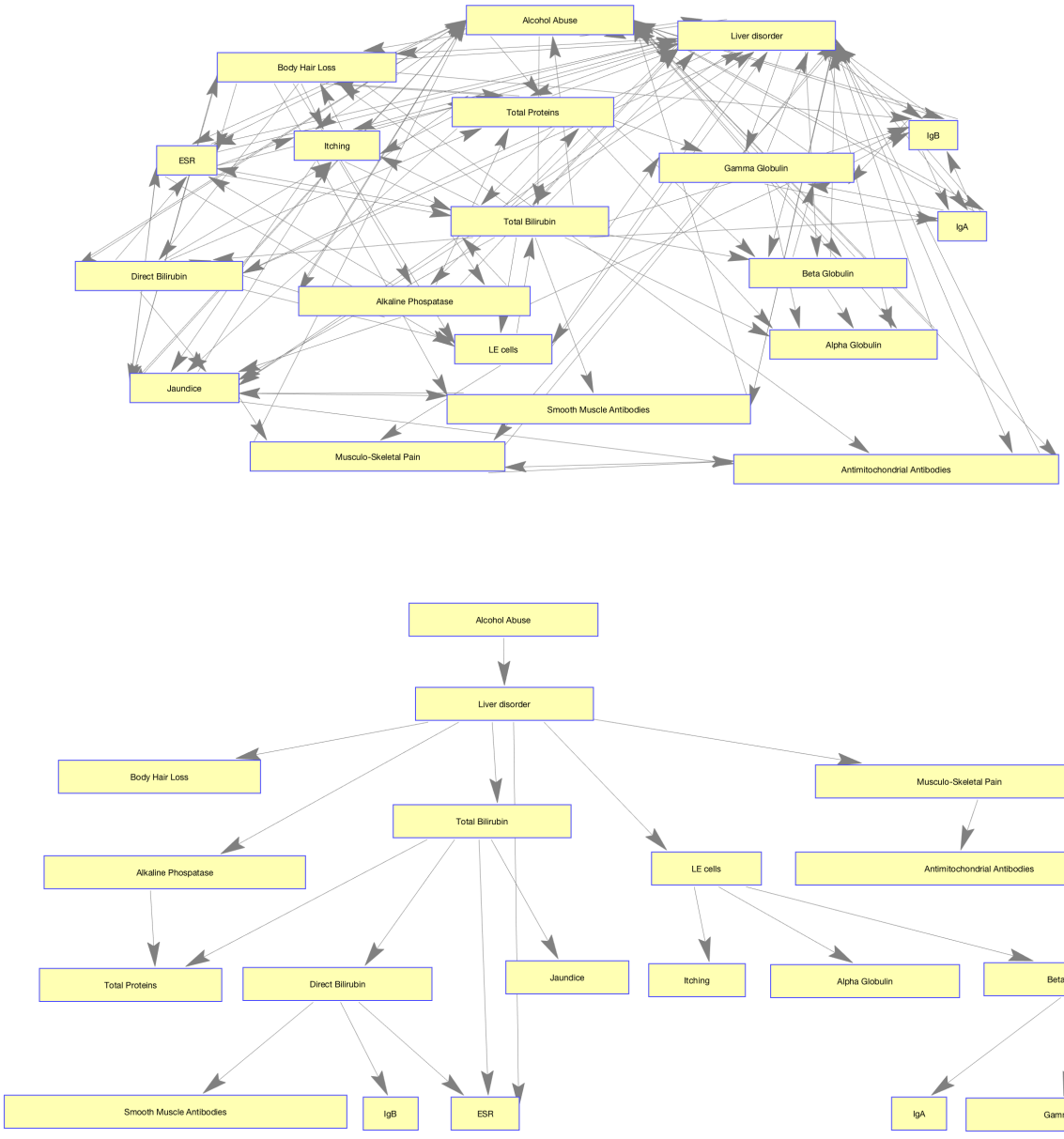



Figure 4: The complete network and the BN identified according to the keyword "liver disorder".

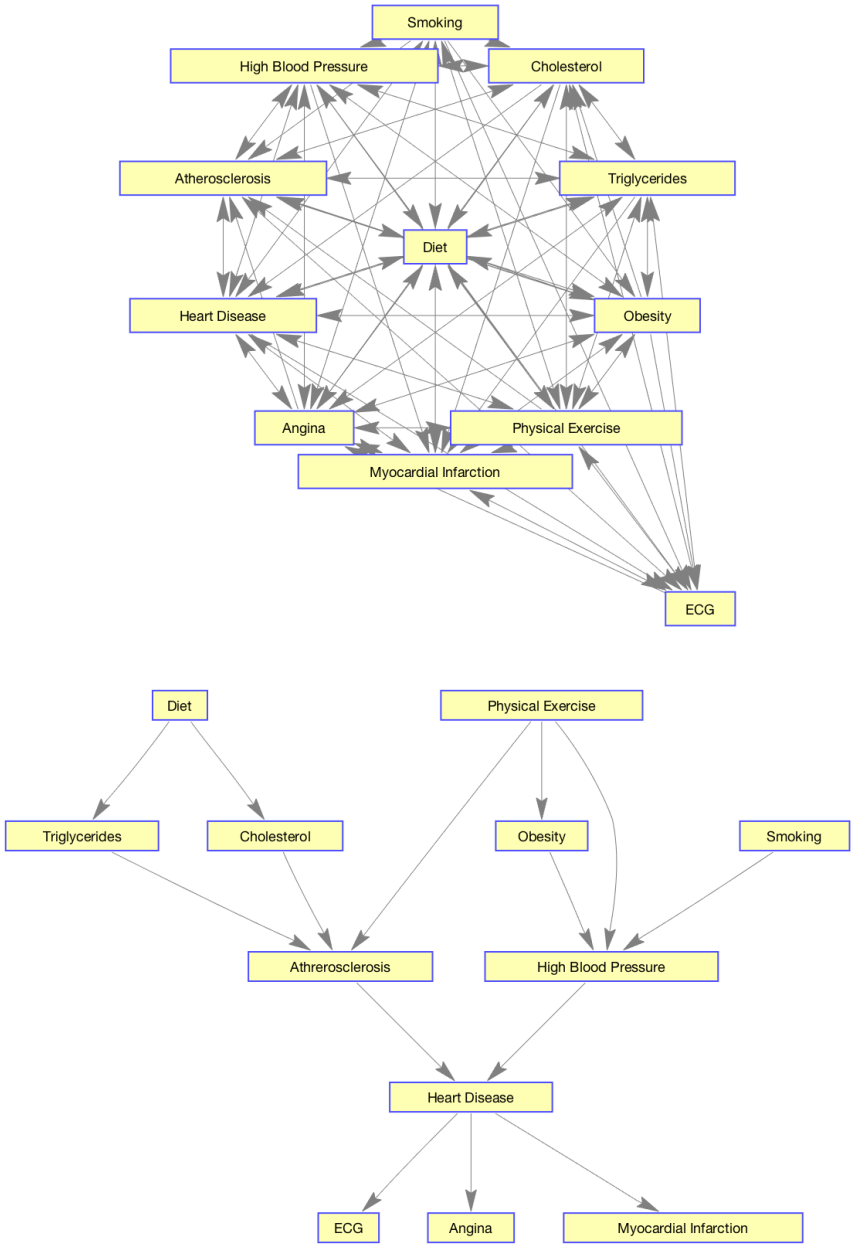
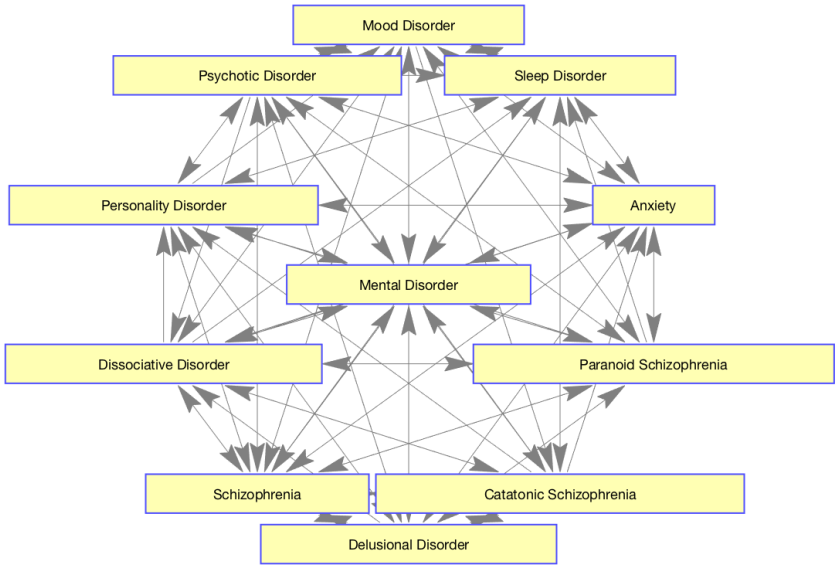



Figure 5: The complete network and the BN identified according to the keyword "heart disease".



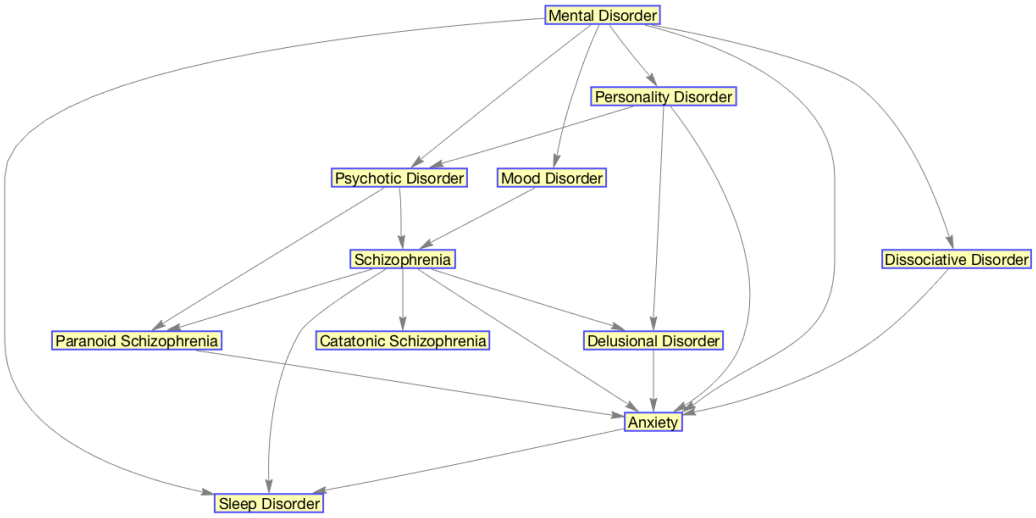




Figure 6: The complete network and the BN identified according to the keyword "mental disorder".

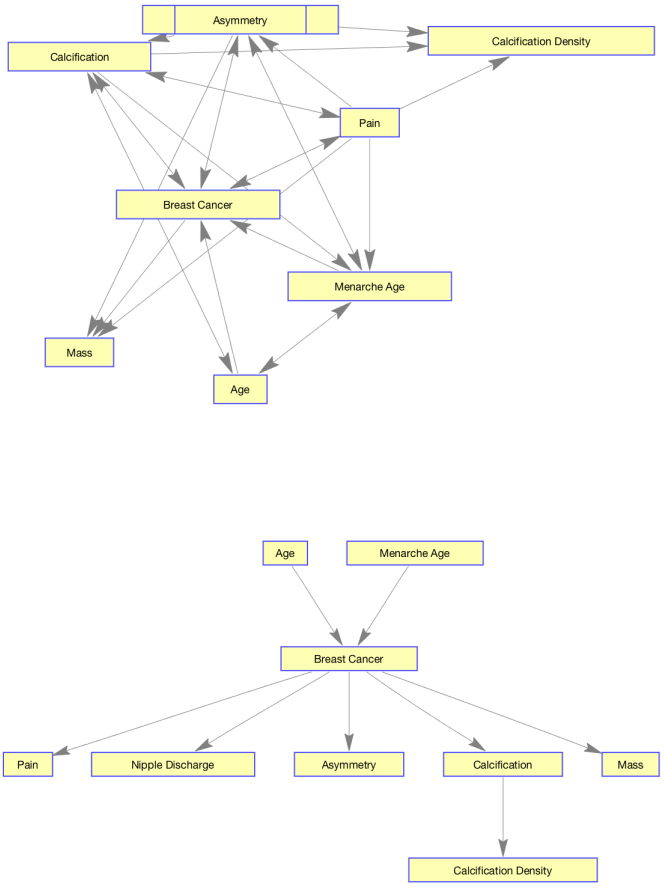



Figure 7: The complete network and the BN identified according to the keyword “breast cancer”.