Towards automatically generating supply chain maps from natural language text

Pascal Wichmann ∗ Alexandra Brintrup ∗ Simon Baker ∗∗ Philip Woodall ∗ Duncan McFarlane ∗

∗ Institute for Manufacturing, University of Cambridge, United Kingdom (e-mail first author: pw351@cam.ac.uk)
∗∗ Department of Computer Science and Technology, University of Cambridge, United Kingdom

Abstract: Supply chains are increasingly global, complex and multi-tiered. Consequently, companies often struggle to maintain complete visibility of their upstream supply network. This poses a problem as visibility of the network is required in order to effectively manage supply chain risk. In this paper, we discuss supply chain mapping as a means of maintaining (structural) visibility of a company’s supply chain, and we derive the requirements for automatically generating supply chain maps from openly available text sources. Early results show that supply chain mapping solutions generated by Natural Language Processing (NLP) could enable companies to a) automatically generate rudimentary supply chain maps, b) verify existing supply chain maps or c) augment existing maps with additional supplier information.

Keywords: supply chain management; supply chain map; supply chain mapping; natural language processing; text mining; supply chain visibility; supply chain mining

1. INTRODUCTION

Supply chains are increasingly global and complex (Christopher and Peck, 2004). Substantial parts of the value creation are outsourced to suppliers who in turn also outsource to sub-tier suppliers themselves. As a result, a company’s performance increasingly depends on the performance of its supply network, and “individual businesses no longer compete as solely autonomous entities, but rather as supply chains” (Lambert and Cooper, 2000). As multi-tiered and geographically distributed supply networks emerge, companies gradually lose visibility which negatively impacts the company’s ability to manage the efficiency, resilience, and sustainability of its supply chain. In particular, supply chain risk management without visibility of the supply network poses a problem to a company while at the same time the emergence of longer, geographically distributed supply chains exposes the company to more and a wider range of risks. Studies show that the share of supply chain disruptions that originate with suppliers further upstream than the direct suppliers can be as high as 50% (KPMG International & The Economist Intelligence Unit, 2013; Business Continuity Institute, 2014) and that suppliers critical to continued operations can be located anywhere in the multi-tiered network and do have to correspond to large sales volumes (Yan et al., 2015). Supply chain mapping is frequently named as the recommended solution to the problem of limited supply chain visibility, yet the actual problem of acquiring the required data in the first place remains unaddressed (Farris, 2010). One of the main reasons for the limited availability of supply chain structure data is the “proprietary nature of each supplier’s relationships with its partners” (Sheffi, 2005).

While data that can readily be used for supply chain mapping is still scarce, we live in an age of an ever increasing availability of massive amounts of data. Vast amounts of data have become abundantly available at low cost via the Web. A large proportion of this data is in natural language form and contains valuable information about buyer-supplier relations.

In this paper, we derive the requirements for automatically generating supply chain maps from natural language text, a process we call “supply chain mining”. This is a first step to enable companies to quickly generate a rudimentary supply network, cross-check an existing model or regularly augment it with additional information.

After introducing supply chains, supply chain visibility as well as NLP (Section 2), we define the supply chain mapping problem and argue why natural language text can be a valuable public data source (Section 3). We then derive the requirements for a supply chain mapping solution using Natural Language Processing methods (Section 4). Subsequently, we test the requirements using a case example from the automotive manufacturer Toyota (Section 5). Finally, we discuss our work and its main limitations as well as propose ideas for future research (Section 6).

2. RELATED WORK

In this section, we offer a brief summary of the related work in the context of supply chain mapping as well as natural language processing.

2.1 Supply chains

A supply chain emerges as a focal company (here also referred to as Original Equipment Manufacturer or OEM)
buys products or services from a supplier to produce their own products. Since supply chains are networks (Lambert and Cooper, 2000), they consist of nodes and directed links of “flows of products, services, finances, and/or information from a source to a customer” (Mentzer et al., 2001). The combination of nodes and links give the network its structural dimensions. The horizontal structure refers to the number of tiers across the supply chain. The vertical structure refers to the number of suppliers or customers represented within each tier (Lambert and Cooper, 2000). The term “upstream” is used to denote the direction towards to original supplier whereas “downstream” refers to the direction towards the ultimate customer.

In academic literature, the term supply chain visibility (also referred to as supply chain transparency) has been defined in various ways. For a comprehensive overview the reader may refer to Goh et al. (2009). Within the scope of this paper, we adopt the broader definition by Barratt and Oke (2007) who define supply chain visibility as “the extent to which actors within the supply chain have access to or share timely information about supply chain operations, other actors and management which they consider as being key or useful to their operations”. Included in the above definition is the knowledge of actors and the network of their dependencies, which we will refer to as the structural supply chain visibility.

Structural supply chain visibility is often limited: A study by Achilles, a provider of supply chain management solutions, claims that “40% of companies who sourced only in the UK, and almost 20% who sourced globally, had no supply chain information beyond their direct suppliers” (Achilles Group, 2013). The reason for limited supply chain visibility is a combination of multiple factors. The main reason is that suppliers have an incentive not to disclose their own supply network to their customers, especially if they run the risk of being cut out as the middleman or losing bargaining power (Sheffi, 2005). The difficulty of obtaining the required data is exacerbated by the fact that supply chains are dynamic (Lambert and Cooper, 2000). The use of tracking technology such as RFID has been explored to increase general supply chain visibility but tracking technologies cannot discover otherwise unknown supply chain participants and their inter-relations.

2.2 Supply chain mapping

Supply chain maps are “a representation of the linkages and members of a supply chain along with some information about the overall nature of the entire map” (Gardner and Cooper, 2003) and aim to address the problem of limited structural supply chain visibility. The purpose of supply chain maps, and hence the scope and level of detail, can vary (Gardner and Cooper, 2003). Their purpose is generally strategic and they range from a geographic vulnerability map which “simply depicts which supplier of what parts are located in each area of the world” (Sheffi, 2005) to maps that show “the flow of parts out of given regions, depicting who is involved and the plants in other parts of the world that are dependent on them” (Sheffi, 2005). Supply chains may or may not depict actual geographical relationships (Gardner and Cooper, 2003).

In this paper, we refer to supply chain mapping as the overall process of creating and maintaining a supply chain map that includes analysing the information needs, acquiring and analysing the information and visualising the results on the required aggregation level.

Gardner and Cooper (2003) provide a comprehensive overview of the visual mapping process. Confronted with the difficulty of obtaining data that spans the complete supply chain, Farris (2010) suggests the use of manually created “macro industry maps” that identify the overall structure of the supply chain at the industry level and then serve as a basis for further more detailed mapping.

The main limiting factor of supply chain mapping is the difficulty of obtaining information about supply chain participants and their inter-relations across multiple tiers. Ideally, the data acquisition process should be automated so that supply chain maps can be easily created and updated without cost-prohibitive manual work.

2.3 Natural language processing (NLP)

NLP is a rapidly developing sub-field of Artificial Intelligence (AI) that specialises in the extraction and manipulation of natural language text or speech (Chowdhury, 2003). Modern NLP methods increasingly rely on Machine Learning. In this work, we focus on Information Extraction (IE), a fundamental task of NLP that aims to automatically extract structured information from unstructured natural text (Cowie and Lehnert, 1996). This structured information is typically used to construct large knowledge bases, relational databases, and ontologies. IE is subdivided into two subtasks: Named Entity Recognition (NER), which is the subtask of locating and classifying instances (text mentions) of entities with pre-defined categories of interest, and Relation Extraction (RE), which is the task of detecting and classifying semantic relationships between named entity mentions (Bach and Badaskar, 2007). Generally used performance metrics for information retrieval and information extraction systems include precision, the share of retrieved documents that are relevant, and recall, the share of all relevant documents that are retrieved.

NLP can be used to generate network structures from text. For example, when it is used to automatically extract taxonomic and non-taxonomic ontologies from text (Maedche and Staab, 2001). Ontologies form a network structure of directed relations which can be visualised in an ontology graph. NLP has also been applied to Twitter data for the purpose of monitoring supply chains (Chae, 2015), but there is no academic work on using NLP for automating the supply chain mapping process.

3. THE SUPPLY CHAIN MAPPING PROBLEM

3.1 Importance of structural supply chain visibility

Structural supply chain visibility is important for a company to manage the resilience, efficiency, and sustainability of its supply chain. Within the scope of this work, we focus on the resilience aspect from the perspective of a company trying to manage risks from its upstream supply chain. Supply chain resilience can be defined as “the adaptive capability of the supply chain to prepare for unexpected

...
events, respond to disruptions, and recover from them by maintaining continuity of operations at the desired level of connectedness and control over structure and function" (Ponomarov and Holcomb, 2009). As the supply chain structure is the network of dependencies that exposes a company to supply chain risks: Basole and Bellamy (2014) examine the link between structural supply chain visibility and risk management and find that “structural visibility into the lower tiers of the supply network has a significant mitigating impact on cascading risks” and that “enhanced visibility is an important and perhaps essential capability for effective supply chain risk identification and mitigation. Supply chain managers must therefore move beyond a simplified dyadic or triadic view to a more holistic approach when developing risk identification and mitigation strategies”. Examples of obscured risk include suppliers depending on the same sub-tier supplier or high-risk supply chain participants on a sub-tier. The network structure also determines how risk events propagate through the network and if they get absorbed or even amplified (Jüttner et al., 2003). An early detection of and response to risk events would require knowledge about which events are relevant to a company’s supply chain. For this, too, knowledge of the supply chain structure is necessary.

### 3.2 The Supply Chain mapping problem

Since the purpose of supply chain maps vary, so do the desired level of detail and the information extraction requirements. We define our objective to be the automated generation of strategic supply chain maps as defined by Gardner and Cooper (2003). For a basic strategic supply chain map, we now define the key elements and the desired attributes of the map.

#### Basic key elements

We argue that, at the most basic level, a strategic supply chain map shall answer the following question for a given point in time: *who supplies whom with what (for which end-product) from where?* We consider the following elements as fundamental:

- *“Who supplies whom?”*: Directed buyer-supplier relationships between organisations are by definition the minimal information required for describing a supply network (nodes and links).
- *“With what?”*: Provided material, parts or services are fundamental information as buyer-supplier relations are specific to what is provided and organisations could supply each other with different parts.
- *“For which end-product?”*: The end-product, for which the provided material, part or service will be used, provides the link to the OEM.
- *“From where?”*: Risk sources often have a geographical context, and the geolocation of a company’s headquarters as a first approximation allows for a rough geographical vulnerability assessment.
- *“When?”*: Knowing when a buyer-supplier relation was existent is considered fundamental because past, but now non-existent, buyer-supplier relations are likely irrelevant.

### 3.3 The problem of generating supply chain maps from text

Publicly available data that could be used for supply chain mapping comes in different forms: From structured data in databases and ontologies to semi-structured and unstructured data, like news reports and forum entries. Data types include images, audio and video files as well as text. There are no publicly available databases or ontologies for supply chain networks yet. Existing ontologies like DBpedia (Auer et al., 2007), however, can provide company ownership structures, company locations and other contextual information. Text data is abundantly available on the Internet and public websites can been indexed by search engines which makes information retrieval time- and cost-efficient. Once published, a text is immediately available and quickly indexed by search engines. The grammar of natural languages provides universal structure for all texts written in that language and changes only slowly. Thus, a solution can be expected to work reliably for years to come once it works. Through the publication of new texts, the dataset gets continuously updated. Given the absence of already existing, publicly available databases or ontologies for supply chain relations, the problem can be now be narrowed down to: *What are the requirements for automating the generation of supply chain maps from publicly available natural language text?*

### 4. NLP APPROACH TO SUPPLY CHAIN MAPPING

#### 4.1 The case for NLP

Natural language can either be processed manually by humans or in an automated manner by software. By definition, software-based methods fall into the domain of NLP. With respect to manual processing, the use of micro task platforms like Amazon’s Mechanical Turk would be the most efficient option. However, this is relatively costly and requires additional quality control mechanisms. Hence, we focus on NLP and consider micro task platforms only for the acquisition of gold standard data to train or evaluate algorithms.
is, therefore, evident that not all desired information will be reported on the Web. This poses multiple challenges for the framework development:

1. A framework needs to be able to cope with incomplete information.
2. Which ‘gold standard’ data does one use for an evaluation if limited data availability is the problem one is trying to solve in the first place? This issue does not directly impact any solution but influences the way a solution can be evaluated, especially regarding recall.
3. Limited data availability can mean that correctly classified data used for training purposes has to be generated manually.

**Imperfect information:** Sources of information especially on the Web are unreliable and can be contradicting. Data quality (including availability) could also differ for different companies and industries. Information needs to be cross-checked, for example via inter-source agreement.

**Ambiguous entity names:** Entity names need to be disambiguated. Brand names and names of the legal entity are often used interchangeably. For example, it is not immediately clear if the sentence “Interstate supplies Mercedes with batteries in the US” really implies a relation between the legal entities “Interstate Batteries” and “Mercedes-Benz USA, LLC”, or “Daimler AG”.

**Abundance of positive information and lack of negative information:** News reports typically contain a ‘positive’ information bias in the sense that they state that one company supplies the other. News reports are less likely to state that a company does not or no longer supply the other. Extracted information might have to be assigned an industry- or part-specific half-life to reduce confidence over time.

**Frequency of occurrence does not indicate criticality:** In some NLP tasks, the frequency of a relation may be used to derive how important that relation is. Similarly, if it is frequently reported that company A supplies company B, we might assume that company A is a major supplier. However, from a risk management perspective, a buyer-supplier relation can be critical even though it does not correspond to a large sales volume (Yan et al., 2015; Simchi-Levi et al., 2014).

**Company- and industry-specific characteristics:** Some of the aforementioned characteristics, such as supply chain dynamics and data availability, are company- or industry-specific. This makes it difficult to predict how well a supply chain can be mapped since the information extraction performance and information quality are affected by the type of company or industry. Furthermore, the language used to describe buyer-supplier relations can be industry-specific.

**Information on different aggregation levels:** Relevant information about supply chains might be provided on different aggregation levels. For instance, an individual buyer-supplier relation may be referenced as “company A supplies company B with part C”. But important information might also be provided on a higher aggregation level, such as “50% of company A’s suppliers are located in the Los Angeles area” or “Product P consists of 100,000 parts...
Fig. 3. Information on different aggregation levels.

sourced from over 1,500 suppliers in Japan, Germany, and China”. The information may be used to assess the performance or validate other results (Figure 3).

5. CASE STUDY

To validate the requirements outlined in Section 4, we implemented a basic solution and applied it in a case study.

5.1 Case study problem statement

The Toyota supply chain has been chosen as our case study because of its global scale and the fact that it has been subject to various studies. The Toyota supply chain has often been reported in the news improving our chances for obtaining relevant public data, and the automotive supply chain appears to be less secretive than other industries. We used an private automotive industry database (Marklines Automotive Information Platform) as our assumed gold standard. The case problem is to map direct and sub-tier suppliers of Toyota by using a basic implementation that meets the derived requirements. The implementation of a basic solution, called Supply Chain Miner, focusses on the basic functionality and, hence, realises aspects of text retrieval, pre-processing, NLP parsing, information extraction, SC mapping and visualisation. The relation extraction is based on pre-defined lexico-syntactic patterns for demonstration purposes.

5.2 Results

Figure 4 depicts the result of dependency parsing which – starting from the verb – links all tokens of the sentence to their dependencies. In this particular example that uses Stanford’s CoreNLP, “Toyota” is the direct object (“dobj”) of the verb “supplies”. Based on these dependencies, extraction patterns can be defined.

Fig. 4. Dependency parsing for an extracted sentence

The aggregation of individual directed buyer-supplier results in a basic network that can be visualised using the attributes from 1. Figure 5 shows the direct Toyota suppliers extracted automatically.

Beyond just direct suppliers, supply chain maps of industries or even across industries can be generated. Figure 6 shows the result of an extended supply chain mapping experiment. At the most basic level, such a map could allow supply chain managers to discover previously unknown sub-tier suppliers or discover that suppliers purchase from the identical sub-tier supplier. The prototypical

implementation does not yet extract provided parts or the end-product but links companies to an ontology to obtain their location.

Fig. 5. Extracted direct Toyota suppliers

To evaluate the performance of the proof-of-concept, two types of performance have to be distinguished:

- **Assessment A**: How well was the supply chain mapped compared to the true supply chain? In this case, unavailable or retrieved but incorrect information will reduce the measured performance. This assessment requires a gold standard dataset to compare against.

- **Assessment B**: Assuming that the provided text is factually correct and only considering the collected text document, how well were relations extracted from the text? This assessment is independent of the quantity and quality of available input data. Furthermore, this assessment can be performed by any human without any further background knowledge. However, because the dataset is imbalanced, for a human, measuring recall is difficult and could only be done by drawing samples.

With regards to the Toyota case study, 12 direct suppliers were extracted based on 26 sentences supporting these relations. Overall, 2598 sentences or sentence fragments were processed.

**Assessment (A)**: When compared with the gold standard and only considering Toyota’s direct suppliers, the measured precision was 8 out of 12 (66.7%). The following four suppliers did not exist in the automotive database: Mazda, BMW, Tesla, Lotus. The collected sentences, however, did state such a buyer-supplier relation. The automotive supplier database contained 3431 buyer-supplier relations where a buyer’s name was “Toyota”. This would result in a recall of 8 over 3431 direct suppliers. However, suppliers in the database are stated on the level of legal entities and, for example, eight of those are part of the “Panasonic” group.
The supply chain maps were based on myopic 1-to-1 relations and actual sub-tier relations could not be automatically inferred. During the experiments, supply networks emerged that span across multiple industries. In one case, the company “Toray” supplied both the automotive as well as the aerospace industry and connected both networks. This is interesting from a supply chain risk management perspective as interdependencies between industries can impact risk exposure. Well-tested lexico-syntactic patterns are a reliable and transparent way of extracting buyer-supplier relations. However, they are not flexible enough and time-consuming to define. Supervised learning using Deep Learning could be a good complementary approach. It is not well-understood how industry and company characteristics impact data availability: We expect to see a strong influence of industry and company characteristics on the availability of input data and, thus, achievable mapping performance.

A proposed framework based on the requirements stated here will be published in a separate paper. This paper will also contain the details of the proof-of-concept implementation of which we only show the results here.

REFERENCES


