Modeling Company Risk and Importance in Supply Graphs

Lucas Carstens[1(✉)], Jochen L. Leidner[1], Krzysztof Szymanski[2], and Blake Howald[3]

[1] Thomson Reuters, Corporate Research and Development, 30 South Colonnade, England E14 5EP, UK
lucas.carstens@thomsonreuters.com
[2] Thomson Reuters, Platform Group, Slaska 23/25, 81-001 Gdynia, Poland
[3] Thomson Reuters, Platform Group, 610 Opperman Drive, Eagan, MN 55123, USA

Abstract. Managing one's supply chain is a key task in the operational risk management for any business. Human procurement officers can manage only a limited number of key suppliers directly, yet global companies often have thousands of suppliers part of a wider ecosystem, which makes overall risk exposure hard to track. To this end, we present an industrial graph database application to account for direct and indirect (transitive) supplier risk and importance, based on a weighted set of measures: criticality, replaceability, centrality and distance. We describe an implementation of our graph-based model as an interactive and visual supply chain risk and importance explorer. Using a supply network (comprised of approximately 98,000 companies and 220,000 relations) induced from textual data by applying text mining techniques to news stories, we investigate whether our scores may function as a proxy for actual supplier importance, which is generally not known, as supply chain relationships are typically closely guarded trade secrets. To our knowledge, this is the largest-scale graph database and analysis on real supply relations reported to date.

Keywords: Supply chain analysis · Graph analysis · Risk analysis · Vulnerability analysis · Linked data · Procurement

1 Introduction

A supply chain is a complex network of interconnected actors that continually exchange goods, with the goal of producing value for all actors in the supply chain. Though supply chains are growing ever more involved, and remain as vital as ever to companies' success, many companies operate with little insight beyond their first tier suppliers and customers. This means that any disruption occurring removed from a company's immediate view risks to be met with little preparedness, and without mitigation strategies in place. To alleviate such risks of being unprepared it is in the interest of companies to increase visibility in supply chains, identifying not only actors they directly interface and exchange

E. Blomqvist et al. (Eds.): ESWC 2017, Part II, LNCS 10250, pp. 18–32, 2017.
DOI: 10.1007/978-3-319-58451-5_2

goods with, but also those residing in subsequent tiers. In addition much of the management of supply chains within companies are founded upon ad hoc methods, relying heavily on human expert knowledge and intuition.

With our work we present a novel approach to investigating the structure of a company's supply chain, based on insights extracted from free text. We represent relations between companies as a graph, where companies are represented as nodes and supply relations as directed edges, pointing from a supplier to a customer (or *consignee*). Not only does this allow us to interpret relations between companies in a formally defined manner, but it additionally provides the opportunity to investigate links between companies beyond their first tier suppliers and customers. More specifically, we use this graph to identify *peers* of a company within its supply chain that are not only particularly relevant, but that are also exposed to certain risks and thus increase the potential for supply chain disruptions. Our graph-based model captures the connectedness of the supplier-consignee supply chain ecosystem in conjunction with the strength of the relationships and the risk exposure of each company entity, which transitively affects potentially large parts of the graph. Specifically, we have developed a solution that is comprised of two APIs, which together provide an aggregate view of peers that are important suppliers to a company, while also being exposed to certain risks. Peers of a company are extracted from a graph database. A pre-specified number of neighbors, from within a pre-specified distance from the node, are extracted and subsequently scored for their importance to the company and their risk. Such a graph model has the potential to serve as the basis for numerous subsequent experiments, including exploring the resilience (see literature review below).

The remainder of this paper is organized as follows. In Sect. 2 we discuss related work at the intersection of two or more of the fields of interest to our work. In Sect. 3 we describe our data sets as well as the construction of a graph database to store and access it. The method of extracting suppliers from the graph database and subsequently scoring their importance and risk is described in Sect. 4. We describe the APIs through which the scoring methods are invoked, as well as the system with which we represent the API output, in Sect. 5. We present an empirical evaluation of the quality of the importance scores in Sect. 6, before concluding the paper in Sect. 7.

2 Related Work

Our work spans a number of fields, touching upon risk and graph analysis, as well as the more nascent area of scientific supply chain analysis, all of which we base on content extracted from the web. Little research has considered the application of both risk and graph analysis to supply chains and modeling. The work of Wagner and Neshat [22], who investigate supply chain risk quantification and mitigation based on graph theory, presents a notable exception. In the remainder of this section we thus review work that resides at the intersection of at least two of the areas we are concerned with. For more specific surveys of the individual

fields refer, for example, to [1] for an overview of graph analysis, to [4] for research on risk and to [20] for a summary of recent advances in supply chain management and procurement.

Supply Graphs. Recent trends in the analysis of supply chains have highlighted the value of representing supply chains as graphs, or networks, rather than as flat structures and relational databases. To this end, Borgatti and colleagues [6] provide an overview of social network analysis, geared towards supply chain research. In the same vein, Kim and colleagues [12] interpret supply chains as networks and apply social network analysis metrics, such as *closeness* or *betweenness centrality*, to evaluate the flow of materials through a supply chains, as well as contractual relationships. Interpreting supply chains as graphs produces wholly new opportunities to investigate structural characteristics and transitive links of complex relations. This is not limited to risk or importance, the focus of our work, but extends well beyond these metrics. For example, Tan and colleagues [19] propose the use graphs to identify innovation potential throughout a network of interlinked companies. Further exploiting graph capabilities, Xu and colleagues [23] describe an evolutionary mechanism that dynamically grows and alters supply networks, reflecting the dynamic nature of supply relations.

Risk and Importance in Supply Chains. Much of the development in assessing risk in supply chains is based on qualitative studies, using expert opinion and case studies. For example, Blome colleagues [5] investigate whether the 2008 financial crisis has had an impact on how risk is managed and, more specifically, whether any of the stages of risk analysis, risk mitigation and risk monitoring have changed. Similarly, Hallikas and colleagues [8] conduct case studies on eleven companies, operating in either the electronics or metal industry, to illustrate challenges that network co-operation brings to risk management. Aqlan and colleagues [3] describe a risk assessment framework that produces risk scores for *suppliers*, *customers*, *manufacturers*, *transportation* and *commodities*. For each stakeholder, experts are consulted to identify the main risk factors. This produces a quantification similar to what we describe in our work, joining impact potential with the risk of this impact actually materializing. Ghadge and colleagues [7] describe a framework comprised of an iterative process to identify, assess and mitigate supply chain risks. They focus on risk assessment, which is comprised of risk modeling and sensitivity analysis, using both a *risk register* and data collected through interviews, company reports, etc. Harland and colleagues [9] describe a *network risk tool* to address the same challenges. The authors focus on risks arising from product and service complexity, outsourcing, globalization and e-business. Based on a set of surveys and focus groups, Juettner [11] seeks to identify and understand business requirements for SCRM from the perspective of professionals working in the field. To structure overarching issues encountered in her analysis into these levels, Juettner first identifies the extent to which organizations already manage risks in their supply chain and then determines critical issues that arise as part of the implementation of risk management. Simchi-Levi, Schmidt and Wei [17] present a dynamic graph

model, which includes recovery time. Unlike our model, their data is obtained from human questionnaires, not automatic text mining.

Risk and Importance in Graphs. The use of *attack graphs* represents one of the more popular approaches to interpreting risk and adverse events in graphs. Attack graphs, as well as *attack trees* are used to model all possible attacks, or *exploits*, on a network. In an early proposal for the application of attack-graphs to the identification of risks in physical networks Phillips and Swiler [15] coin *network-vulnerability analysis*. In a more recent development, Alhomidi and Reed [2] use a *genetic algorithm (GA)* [13] to model a large number of possible *paths* in attack graphs, where each path connects the source of an attack on a network to the target of the attack. In each path, nodes are assigned with a probability that represents the likelihood of the node being exploited by an attacker, as well as an expected loss, accrued when a node is indeed attacked. In an adoption of attack graphs, Poolsappasit and colleagues [16] propose a framework for dynamically managing security risks called *Bayesian attack graphs (BAG)*. The overall risk of each possible path in an attack graph is calculated as a product of the attack success likelihoods and the value of the expected loss incurred. Based on data for 371 banks that failed during the 2008 financial crisis, Huang and colleagues [10] study the systemic risk of financial systems. To do so they propose a cascading failure model to describe the risk propagation process during crises. A bi-partite banking network model is proposed, where one type of node represents banks and another represents assets held by banks. The resulting graph is *shocked* by decreasing the *total market value* of an asset, leading to a decrease in value for every bank that holds the affected asset. Stergiopoulos and colleagues [18] extend the notion of cascading failure models to include *graph centrality* measures to help identify the nodes most critical in identifying and mitigating failures. Graph centrality here is used as a proxy for identifying the *most important* nodes within a graph so that any risk mitigation strategy can be based on both the importance of nodes and their susceptibility to failure, in general. None of the work surveyed above provides a larger-scale supply chain graph model, which can be used for the analysis of and experimentation with supply relation scoring methods. Below we present such a model, as well as its implementation.

3 Building a Supply Graph

Our analysis of supply relations between companies is based on a graph database, where the nodes represent companies and edges signify directed supply relations, pointing from a supplier to a customer. Supplier/customer pairs are extracted automatically from specifically news articles. Each node in the graph is assigned a set of attributes, namely (i) their business sector, (ii) a credit risk score, (iii) the company name and (iv) a closeness centrality score. We describe the data extraction process, as well as the node attributes, in detail below, for two separate supply graphs. On the one hand we conduct experiments, as described in Sect. 6, on the full graph (SPR^{+}), with all its attributes. On the other hand,

Table 1. Dataset: summary statistics.

Number of nodes	98,402
Number of vertices	217,188
Average path length	6.614
Average degree	4.414
Average closeness centrality	0.225

necessitated by the proprietary nature of this data, we make available a second dataset (SPR^{-}) for research purposes[1]; in it, company names and business sectors are anonymized. Table 1 summarizes the main characteristics of the data.

3.1 Supply Relations Data

Both datasets, $SPR^{+/-}$ are comprised of supply relations between two companies, where each individual relation and the companies involved are automatically extracted form text snippets. While we describe a static snapshot of the dataset for our experiments, an underlying *RDF triple store* of supply relations is continually updated, both to add new relations and to remove those considered out of date. A snippet corresponds to a sentence, extracted either from a news article or a *Security and Exchange Commission (SEC)* filing. A logistic regression model was trained on a set of 45,000 snippets, while the test set was comprised of 20,000 snippets. The training and test data were aggregated using the following procedure:

1. Identify companies in a document, using *Calais*[2];
2. split documents into sentences;
3. choose candidate sentences that contain two companies, as well as one or more of a pre-specified set of patterns; and
4. using *Mechanical Turk*, label companies in the candidate sentences as suppliers, customers, or neither.

Patterns are based on a set of indicative *n-grams*, as well as variations of these n-grams to catch terms such as *powered by*, *contracts with*, etc. Each candidate sentence has been labeled by two separate *Turkers* and any disagreement was addressed by presenting the instance to a third annotator. The regression model has been tuned to yield high precision, focusing on the extraction of high quality evidence sentences, while relying on the fact that eventually, highly indicative sentences will be introduced into the dataset. The classifier produced an F_1-score $F_1 = 0.57$ (with $precision = 0.76$ and $recall = 0.46$) on the test set (note that a random baseline classifier would achieve an accuracy of 0.5). Data is stored in an RDF triple store from which we can extract a subset or, to

[1] see http://bit.ly/TRSupplyChainRisk.
[2] http://www.opencalais.com/.

populate our graph database, the entire set using *Sparql* queries. This triple store is continually updated to add additional relations found in unseen text; multiple patterns producing the same supplier-consignee pairs are aggregated to a single triple. Each triple has a confidence score assigned, based on the classifier output as well as the amount of examples found for a specific relation.

3.2 Company Attributes

To score a company according to its importance as a supplier to a customer, as well as the risk it is exposed to, we assign a set of attributes to each company, in addition to its name for identification purposes. The importance of a company is then determined based on how a supplier's attributes compares to those of the customer, as well as both their position in the overall graph.

Business Sector. Each company in the supply graph is labeled with the *business sector* it operates in. We use the *Thomson Reuters Business Classification (TRBC)*[3] scheme for this purpose, a widely used industry standard. The TRBC scheme offers classification of companies at various levels of abstraction, i.e. economic sectors, the most abstract level, business sectors, industry groups, industries, and activities. To strike a compromise between informativeness and the ability to group various companies we label companies with their business sector, meaning that we distinguish between 28 different labels, such as *Renewable Energy*, *Industrial goods*, etc.

Credit Risk. To identify the risk companies are exposed to we score them according to a credit risk measure. This score broadly signifies the likelihood of a company defaulting on one or more of their debt obligations within a year. A score between zero and 100 is used to signify the likelihood, where a lower score represents a higher likelihood of default.

Closeness Centrality. One of the aims of our importance scoring is to incorporate both attributes of individual companies and those formalizing a company's role within a larger graph of companies. To this end we have chosen to score each node in the graph according to its *closeness centrality*. Closeness centrality measures a node's centrality in a graph as the sum of the length of the shortest paths between the node and all other nodes in the graph. This sum is usually normalized by division with the total node count N (minus one so as not to count the node itself) to represent the average length of the shortest paths, or distance $d(y, x)$, giving

$$C(x) = \frac{N - 1}{\sum_y d(y, x)}, \tag{1}$$

3.3 Database

The data described above is initially extracted from the RDF store and represented as separate node and edge tables, which are, in turn, used to populate a

[3] http://financial.thomsonreuters.com/en/products/data-analytics/market-data/indices/trbc-indices.html.

graph database, implemented using Neo4j. While the original dataset described here contains proprietary data and can hence not be published, we make available an anonymized version of the node and edge tables. To interact with the database we use *Cypher*, the graph query language developed as part of Neo4j. To load the node table we call the following command

```
USING PERIODIC COMMIT
LOAD CSV WITH HEADERS FROM 'file:///file_path/nodes.csv' as line
WITH line MERGE (ID:permID {name: TOINT(line.permID)})
SET ID.trbc = line.TRBC, ID.centrality = line.centrality,
ID.company_name = line.company_name,
ID.ccgr = line.CCGR;
```

and, to load the edge table, we call

```
USING PERIODIC COMMIT
LOAD CSV WITH HEADERS FROM 'file:///file_path/rels.csv' as line_b
MATCH (sup:permID {name: TOINT(line_b.supplier)})
WITH sup, line_b MATCH (cus:permID {name: TOINT(line_b.customer)})
MERGE (sup)-[:supplies]->(cus)
```

This populates a previously initialized Neo4j instance, which can then be queried. In our case we want to evaluate the importance and risk of a pre-defined number of suppliers to a specific customer. To do so, we need to identify a single node within the database, i.e. the customer, and query for neighbours whose directionality points towards that node, i.e. the company's suppliers. Depending on the setting of the query we may do this recursively to not only retrieve direct suppliers, but suppliers of suppliers, also. We generally refer to direct suppliers as *first-tier* suppliers, to suppliers of suppliers as *second-tier* suppliers, and so forth. The below retrieves up to 1,000 first- and second-tier suppliers of the node *0123456789*.

```
MATCH (n:permID {name: 0123456789}),
p=shortestPath((x)-[:supplies*1..2]->(n))
WITH LENGTH(p) AS lp, x LIMIT 1001
RETURN
x.name, x.trbc, x.ccgr, x.centrality, x.company_name, lp;
```

Note that we set the limit to 1,001 because the node we are searching for is included in the limit, as well. While we have used Neo4j as the database of choice we opted to run graph analyses using *Gephi*[4]. On the one hand we have used Gephi to calculate closeness centrality scores for nodes in the graph, as described above. On the other hand Gephi provided a natural interface to run initial analyses on the graph to determine its overall structure. This includes calculating the measures reported in Table 1.

[4] https://gephi.org.

3.4 World Input-Output Database (WIOD)

The *World Input-Output Database (WIOD)* [21] provides data on the distribution of supply activities between business sectors.[5] We use this data as part of our importance calculation, where we compare the business sector the supplier operates in with the business sector the customer resides in. The WIOD allows us to deduce whether these two industries have a strong relation, in terms of relative volume exchanged between the business sectors, compared to other business sector combinations. The WIOD is comprised of supply data between a total of 43 countries and compares business sectors based on the *International Standard Industrial Classification (ISIC)*. Data is collected for the period between 2000 and 2014. Since for our work we have assigned TRBC codes to the companies that comprise the supply graph we use a mapping between TRBC and ISIC codes that has been created internally to align the WIOD with our data.

4 Scoring Method

The graph database described in the previous section facilitates the analysis of supply relations between companies within the context of a larger network. In this section we describe how we use the graph database to identify relevant suppliers of a customer through multiple tiers of the supply graph and score them according to two metrics, (1) importance and (2) risk. Importance, described in detail in Sect. 4.1, scores suppliers of a company based on a combination of metrics, incorporating both the structure of the graph and the supplier's position in it, and attributes of the supplier itself. With it we aim to quantify the adverse impact that a disruption to the supply from a specific supplier would have on a specific customer, where a high importance, i.e. a score close to 1, reflects a high potential adverse impact. Risk, described in detail in Sect. 4.2, is scored according the *credit risk scores* assigned to each company in the graph.

4.1 Scoring Supply Chain Importance

We calculate importance scores $I = (i_0, ..., i_n)$ for suppliers $S = (s_0, ..., s_n)$, each represented by a node in a graph, retrieved from the graph database. The nodes are retrieved in relation to node c, representing a customer, representing the n companies closest to c. Each importance score i_m is an aggregate of four measures:

a. *Criticality*: The proportion of goods the business sector of q receives from the business sector of i_m (based on WIOD data, see Sect. 3.4);

$$a = \frac{Criticality}{m}; \tag{2}$$

where m is a normalization constant $m = 34.27$. The constant represents the strongest tie between any two industries in the WIOD dataset.

[5] http://www.wiod.org/home.

b. *Replaceability*: The sum of how many s operate in the same business sector as s_m (based on TRBC codes):

$$b = 1 - \left(\frac{Replaceability}{n-1} \right); \tag{3}$$

c. *Centrality*: a metric of the importance of s_m to the overall graph (we can use *closeness centrality*, for instance);

$$c = Centrality \tag{4}$$

d. *Distance*: the (step-)distance between s_m and c.

$$d = \frac{n - Distance}{n-1} \tag{5}$$

We then aggregate i_m as follows:

$$i_m = \frac{\left(\frac{a+b+c+d}{4}\right)}{max \in I} \tag{6}$$

The above operations normalize all individual scores to a value in the range $[0; 1]$. The scores are also normalized so that a value closer to one reflects a higher importance. We also normalize i_m by dividing its result by the maximum score of all i, so that the *most important* node always has a score of one and all other nodes are scored in relation to it. We catch the fringe case that yields division by zero programmatically (where $n = 1$), in which case we can simply set $i_0 \leftarrow 1$.

4.2 Scoring Supply Chain Risk

The second metric according to which we score the suppliers is credit risk. Akin to importance we calculate risk scores $R = (r_0, ..., r_n)$ for suppliers $S = (s_0, ..., s_n)$ of customer c. Each risk score r_m is based on a single attribute of a node, namely one of two scores, (1) *Credit Combined Global Rank* or (2) *Private Company SmartRatios Global Rank*. Score (1) is assigned to public companies, while score (2) is used for private companies. Both scores are extracted from proprietary Thomson Reuters solutions. The coverage of risk scores for companies that comprise the supply chain agreement dataset is roughly 26%. To cover the gaps we heuristically determined risk scores for companies without a risk score. In a first step we grouped companies based on their business sector, using TRBC codes, and calculated the average risk for each business sector, using the available scores. Companies with missing risk scores were then assigned the average risk score according to their TRBC code. Once each node in the graph had a risk score assigned we normalized the score so that its range is between zero and one, and a higher score represents a higher risk.

5 System Description

The implementation of our scoring methodology is comprised of two components. On the one hand we have implemented two APIs to expose scoring algorithms, one each to execute the importance and risk scoring for the suppliers retrieved form the graph database and returning the results as *json* files. On the other hand we set up a profile for an existing interface to dynamically visualize the results.

5.1 Application Programming Interfaces

The scoring algorithms described in the previous section are accessed through separate Application Programming Interfaces (APIs), each of which accept as arguments the following three parameters; *company ID*, *node count* and *depth count*. In the original dataset we use *permIDs*[6] as company IDs, which have been replaced by random ten-digit IDs in the public dataset.

The node count determines how many neighbours of the node representing the company ID are retrieved from the graph, while the depth count determines from how many tiers we retrieve neighbours. The system accepts two API calls, one to score supply chain importance and another to score risk. Each of the two APIs returns a *json* file with the following format:

```
{
    "dimensionName": "Supply chain importance",
    "peers":[
    {
            "eid": "0022446688",
            "name": "c",
            "score": 1,
            "baseEntity": true
    },{
            "eid": "8800224466",
            "name": "s_0",
            "score": i_0,
            "baseEntity": false
    },{
            "eid": "6688002244",
            "name": "s_1",
            "score": i_1,
            "baseEntity": false
    },...{
            "eid": "4466880022",
            "name": "s_n",
            "score": i_n,
            "baseEntity": false
    }]
}
```

[6] https://permid.org/.

The output JSON file is comprised of the header and two types of blocks. The header identifies which dimension the scores in the JSON file represent. In the example this is the importance score. The first block following the header represents the input entity, i.e. the customer passed to the API. The *baseEntity* label is set to true to represent this and the score is set to a placeholder value of one. Each subsequent block represents a supplier of the *baseEntity*, which may be a supplier at any tier, depending on the parameter settings. Each block is comprised of the ID (*eid*), uniquely identifying the company, the company's name, its importance score and the *baseEntity* flag set to $false$. The output of the risk scoring API produces the same structure, the only difference being the dimension name in the header.

5.2 Interface

The importance and risk calculations, exposed through the APIs described above, are queried, and results visualized, using an internal application called *Jersey*. Jersey provides an interface with functionalities to search for an entity, here companies, and retrieve and visualize the entity's peers. Once a user submits a query, Jersey requests data through our APIs and renders the returned json files as shown in Fig. 1 or 2. Figure 1 shows importance as a single slider, aggregated as described in Sect. 4.1. Figure 2, on the other hand, returns the components of our importance score individually. In both cases the user can use the sliders to adjust the weighting of the individual scores, depending on individual preferences, with the second view allowing more granular weighting. Additionally Jersey offers options to use different weighting mechanisms and render more or fewer results.

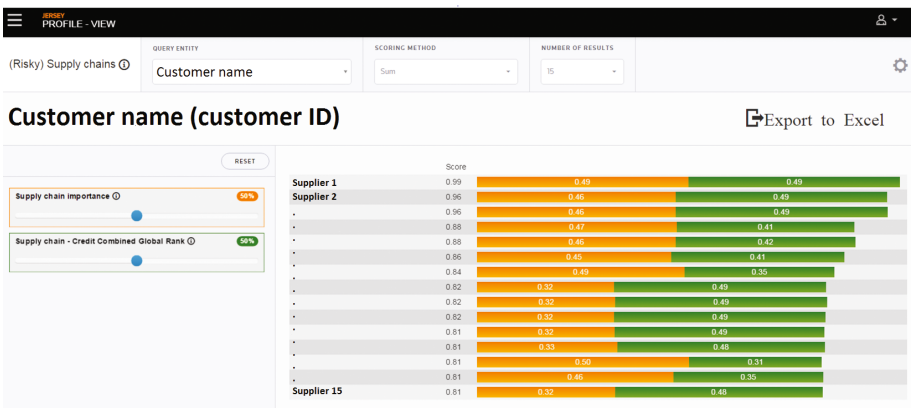

Fig. 1. Our application's web interface.

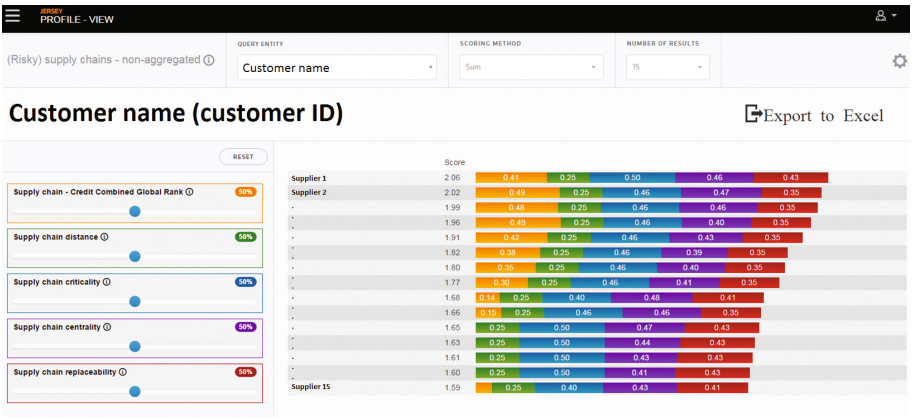

Fig. 2. Importance scores can be broken down into components.

6 Experiments

The risk score we have used is an industry standard measure of credit risk and we have thus focused our evaluation on determining whether the importance score, which we have developed as part of our work, produces a representative measure of the actual importance of a supplier to its customer. As described in Sect. 3.1, the supply chain agreements between two companies that comprise the graph are extracted from natural language text. To evaluate the veracity of how we score these relations on the importance we use the snippets from which the relations have been extracted. To build an evaluation dataset we have done the following:

1. Select 200 snippets that contain correctly classified supply relations.
2. Identify the corresponding supply agreement and score their importance using our application.
3. Select the 20 highest scoring, as well as the 20 lowest scoring snippets
4. Randomly match a high scoring snippet with a low scoring snippet to create a set of 200 pairs of snippets

Two annotators were then asked to the snippet in a pair that described a more *important* supply relation. To determine the veracity of the generated importance scores we then measured correlation between the scores and human judgments. The annotations produced rather low inter-annotator agreement, with a *Cohen's kappa* of $\kappa = 0.2499$. As expected, then, a *Pearson correlation coefficient* produces a similarly low score of $p = 0.1451$.

Discussion. Considering the low kappa score it appears that deciding the importance based solely on comparing snippets may not be a viable approach to judging the veracity of our importance score. The random matching of snippets without pre-selection may have also hurt our evaluation. Note, however, that we

do not calculate importance based on judging the snippets and thus low correlation need not mean that our importance scores do not reflect reality. What we can conclude, based on the low kappa value, is that human judgement of importance based on short snippets does not seem viable and, based on the low p value, that we cannot confirm the veracity of our importance scores. A further limitation of our approach is that, because no fine-grained data at the level of each Other challenges that may have hurt performance are the data gaps described in Sect. 4.2 and the simple, equal weighting that produces the aggregate importance score, as described in Sect. 4.1 individual transaction between a supplier and consignee was available for this study, we had to estimate the connections, somewhat crudely, taking sector-to-sector flows as proxies. This will no doubt have had adversarial effects, and it is hard to quantify them. Nevertheless, we believe that having a quantitative model is valuable, and if and when more granular data becomes available, more refined models can be compared to our model to demonstrate their merit.

7 Summary, Conclusions and Future Work

With this paper we have presented a novel approach to evaluating suppliers of companies according to their importance and risk. We have developed a dataset comprised of roughly 98,000 companies and 220,000 supply relations, which we represent as a directed graph. Using a combination of a company's attributes and their position within the graph we determine their importance as suppliers to a specific company. This has allowed us not only to develop a principled representation of a supply graph, it also allows us to investigate supply beyond the first tier suppliers of a company.

One of the main shortcomings of our work is the limited insight we can gain from the evaluation. In future work we will need to identify new ways of evaluating the importance scores, based on which we can develop new weighting algorithms. Beyond this immediate concern the following three main avenues for further research present themselves. First, extending the scope of risks we measure will add to the informativeness of the overall scoring. For example, we may include evidence from text-mining based risk analysis approaches as described in [14]. Secondly, an extension of our approach to languages other than English will vastly expand the solution's usefulness from a practical application point of view. Whether such an extension ought to be based on machine translation or purpose built models for each language in scope remains to be seen. Finally, we believe that developing a mechanism by which we can either learn the weights of the individual importance scores or determine them through a grid search, rather than simply weighting them equally, may further enhance the quality of the importance scores assigned to companies.

Acknowledgments. We would like to thank Khalid Al-Kofahi and the CTO office for supporting this work and thank Giuseppe Saltini, Shai Hertz, Yoni Mataraso and Geoffrey Horrell for discussions and data.

References

1. Aggarwal, C.C.: An introduction to social network data analytics. In: Aggarwal, C.C. (ed.) Social Network Data Analytics, pp. 1–15. Springer, Heidelberg (2011)
2. Alhomidi, M., Reed, M.: Attack graph-based risk assessment and optimisation approach. Int. J. Netw. Secur. Appl. **6**(3), 31 (2014)
3. Aqlan, F., Lam, S.S.: A fuzzy-based integrated framework for supply chain risk assessment. Int. J. Prod. Econ. **161**, 54–63 (2015)
4. Bisias, D., Flood, M.D., Lo, A.W., Valavanis, S.: A survey of systemic risk analytics. US Department of Treasury, Office of Financial Research 0001 (2012)
5. Blome, C., Schoenherr, T.: Supply chain risk management in financial crises - a multiple case-study approach. Int. J. Prod. Econ. **134**(1), 43–57 (2011)
6. Borgatti, S.P., Li, X.: On social network analysis in a supply chain context. J. Supply Chain Manage. **45**(2), 5–22 (2009)
7. Ghadge, A., Dani, S., Chester, M., Kalawsky, R.: A systems approach for modelling supply chain risks. Supply Chain Manage. Int. J. **18**(5), 523–538 (2013)
8. Hallikas, J., Karvonen, I., Pulkkinen, U., Virolainen, V.-M., Tuominen, M.: Risk management processes in supplier networks. Int. J. Prod. Econ. **90**(1), 47–58 (2004)
9. Harland, C., Brenchley, R., Walker, H.: Risk in supply networks. J. Purch. Supply Manage. **9**(2), 51–62 (2003)
10. Huang, X., Vodenska, I., Havlin, S., Stanley, H.E.: Cascading failures in bi-partite graphs: model for systemic risk propagation. Sci. Rep. 3, Article no: 1219 (2013). doi:10.1038/srep01219
11. Jüttner, U.: Supply chain risk management: understanding the business requirements from a practitioner perspective. Int. J. Logist. Manage. **16**(1), 120–141 (2005)
12. Kim, Y., Choi, T.Y., Yan, T., Dooley, K.: Structural investigation of supply networks: a social network analysis approach. J. Oper. Manage. **29**(3), 194–211 (2011)
13. Mitchell, M.: An Introduction to Genetic Algorithms. MIT Press, Cambridge (1998)
14. Nugent, T., Leidner, J.L.: Risk mining: company-risk identification from unstructured sources. In: IEEE International Conference on Data Mining, ICDM, pp. 1308–1311 (2016)
15. Phillips, C.A., Swiler, L.P.: A graph-based system for network-vulnerability analysis. In: Proceedings of the 1998 Workshop on New Security Paradigms, Charlottsville, VA, USA, September 22–25, 1998, pp. 71–79 (1998)
16. Poolsappasit, N., Dewri, R., Ray, I.: Dynamic security risk management using Bayesian attack graphs. IEEE Trans. Dependable Sec. Comp. **9**(1), 61–74 (2012)
17. Simchi-Levi, D., Schmidt, W., Wei, Y.: From superstroms to factory fires: managing unpredictable supply chain disruptions. Harv. Bus. Rev. **92**(1), 96–100 (2014)
18. Stergiopoulos, G., Kotzanikolaou, P., Theocharidou, M., Gritzalis, D.: Risk mitigation strategies for critical infrastructures based on graph centrality analysis. IJCIP **10**, 34–44 (2015)
19. Tan, K.H., Zhan, Y., Ji, G., Ye, F., Chang, C.: Harvesting big data to enhance supply chain innovation capabilities: an analytic infrastructure based on deduction graph. Int. J. Prod. Econ. **165**, 223–233 (2015)
20. Tayur, S., Ganeshan, R., Magazine, M.: Quantitative Models for Supply Chain Management, vol. 17. Springer, Heidelberg (2012)
21. Timmer, M.P., Dietzenbacher, E., Los, B., Stehrer, R., Vries, G.J.: An illustrated user guide to the world input-output database: the case of global automotive production. Rev. Int. Econ. **23**(3), 575–605 (2015)

22. Wagner, S.M., Neshat, N.: Assessing the vulnerability of supply chains using graph theory. Int. J. Prod. Econ. **126**(1), 121–129 (2010)
23. Xu, N.-R., Liu, J.-B., Li, D.-X., and Wang, J.: Research on evolutionary mechanism of agile supply chain network via complex network theory. In: Mathematical Problems in Engineering 2016 (2016)