Reduced Topologically Real-World Networks: a Big-Data Approach

Marcello Trovati
Department of Computing and Mathematics, University of Derby, UK
M.Trovati@derby.ac.uk

Abstract — The topological and dynamical properties of real-world networks have attracted extensive research from a variety of multi-disciplinary fields. They, in fact, model typically big datasets which pose interesting challenges, due to their intrinsic size and complex interactions, as well as the dependencies between their different sub-parts. Therefore, defining networks based on such properties, is unlikely to produce usable information due to their complexity and the data inconsistencies which they typically contain. In this paper, we discuss the evaluation of a method as part of ongoing research which aims to mine data to assess whether their associated networks exhibit properties comparable to well-known structures, namely scale-free, small world and random networks. For this, we will use a large dataset containing information on the seismologic activity recorded by the European-Mediterranean Seismological Centre. We will show that it provides an accurate, agile, and scalable tool to extract useful information. This further motivates our effort to produce a big data analytics tool which will focus on obtaining in-depth intelligence from both structured and unstructured big datasets. This will ultimately lead to a better understanding and prediction of the properties of the system(s) they model.
Keywords: Knowledge discovery; Networks; Information extraction; Data analytics; Text Mining; Seismological data

Introduction

The majority of contemporary scientific advancements have been based on the ability to identify specific properties of data, and provide both analytical and predictive capabilities. Furthermore, with the increasing availability of big-data sets, new challenges, as well as opportunities have risen which are at the very core of Big Data research.

In particular, data come in a variety of types, forms, and size, which makes the way we extract and assess information a crucial step in gathering intelligence. However, big data-sets need to be suitably manipulated and assessed to ensure they can be effectively analysed.

In this paper we introduce a novel method to topologically reduce networks created by the elements of data-sets, and their mutual relationships. This provides a tool to superimpose networks on top of real-world data to describe their main properties, whilst providing a computationally efficient method.

Network theory has been developed since the birth of discrete and combinatorial mathematics (Bollobas, 1998)which, broadly speaking, aims to describe and represent relations, referred to as *edges*, between objects, or *nodes.* In particular, it has a huge set of applications within a variety of multi-disciplinary research fields, including applied mathematics, psychology, biomedical research, computer science, to name but a few (Dingli, et al., 2012) .

Formally, networks are defined as a collection of nodes, called the *node set* $V=\{vi\}i=1n$, which are connected as specified by the *edge set* $E=\{eij\}i\neq j=1n$ (Albert, et al. 2002)..

Although networks are based on relatively simple mathematical concepts, their general properties exhibit powerful features that can be applied to model complex scenarios (Trovati, et al., 2014)

Data often consist of elements, which could be numeric values, physical entities, or general semantic concepts, which are linked by relationships. Despite its intrinsic vagueness, this can be effectively described by using networks, even though populating the edge and node sets is typically a complex task. In fact, extracting the relevant information can be challenging especially when addressing unstructured data-sets. Furthermore, when size plays a crucial role, such as in Big Data, such extraction can be even more difficult to carry out effectively. Therefore, there are several methods to generate networks from data, which can be, in turn, investigated according to the overall features of such networks.

One of the most important parts in this investigation is to determine the topological structure of a network to allow a complete mathematical and statistical investigation of the data set(s) associated with it.

Network analysis techniques have been extensively investigated and the use and applications of network data has been proposed previously in a wide range of real-world complex settings (Akoumianakis, et al., 2012) Zelenkauskaite A, Bessis N, Sotiriadis S and Asimakopoulou E. (2012). Interconnectedness of Complex Systems of Internet of Things through Social Network Analysis for Disaster Management. Proceedings of INCoS 2012: 503-508 (Zelenkauskaite, et al., 2012). In general, it has been found that the majority of network analyses ignore the network itself that it is the actual focus of this work.

Networks are relatively simple to define based on suitably processed data sets. In fact, via data and text mining techniques, it is possible to isolate semantic objects, such as physical, as well as conceptual entities, along with their mutual relationships determined by hierarchical properties of the corresponding data sets.

In this paper, the idea of reducing the topology of a network determined by pre-processed data focuses on its complexity, rather than on its structure in terms of edges and nodes. In other words, we are proposing a method to determine which degree distribution best describes a real-world network, rather than pruning it to decrease its size. Our main goal is to determine which rule, if any at all, can describe the structure of a real-world network. In particular, we, aim to provide a complete toolbox which facilitates intelligence extraction from big data-sets. In particular, this will enable the definition of networks lying on an intermediate layer, which is used to efficiently identify and classify big data. As part of our evaluation, we will analyse the network (and sub-networks) associated with a large dataset containing information on the seismologic activity recorded by the European-Mediterranean Seismological Centre (Zelenkauskaite, et al., 2012) . In particular, we will show that it exhibits a scale-free structure, which indicates the likelihood of a non-random set of events. Furthermore, this also suggests the existence a co-occurrence relationships among the events corresponding to the nodes.

The rest of the paper is organised as follows: in Section Theoretical Background we describe the main features of the networks considered in the paper. Section Big Data discusses the relevance of Big Data and its properties, while Section Description of our Approach focuses on the description and implementation of the main algorithms. Finally, Section Evaluation discusses the evaluation we have carried out, and Section Conclusions and Future Steps concludes and prompts future directions of our work.

Theoretical Background

In this section the main theoretical concepts of network theory are discussed, which are exploited and applied as part of our method.

Random Networks

Random networks have been extensively studied since their introduction (Erdȍs, et al., 1960), when it was realised this could describe the complexity of real-world systems that could not be captured by ordinary deterministic networks. The general definition of such networks is rather simple: we start with n nodes which are mutually connected with an independent probability p. Over the last few decades, more probabilistic models have been defined to extend the above definition.

Random networks are characterised by a probability distribution p, which governs the existence of the edges between any two nodes. In particular, the probability p is linked to the fraction pk of nodes with degree k as follows:

$$pk \approx zke\text{-}zk!$$

where $z=n\text{-}1p$, and n is the number of nodes.

However, it is certainly legitimate to wonder how realistic such type of model really is. If we consider a general social network, a purely random approach would entail that any new member would connect and interact at random. Over time all the new members will spread uniformly over the network. However, one of the main criticisms is that individuals are more likely to interact with socially active people, or in other words, nodes of social networks with a higher degree. This is one of the reasons of the introduction of *scale-free* networks as discussed in the next section. On the other hand, if a network does exhibit a purely random behaviour, it may suggest that the objects that correspond to the nodes, do not have any co-occurrence rule that may indicate an influence between them.

Scale-Free Networks

Numerous physical and, more generally, real-world systems exhibit properties that can be described as scale-free networks, such as biological and bio-medical systems, as well as social networks (Humphries, et al., 2008). . The introduction of this type of network is relatively recent, since network research mainly focused on random networks (Bollobas 1998). In particular, Albert et al. (Albert, et al. 1999). whilst investigating the properties of webpages and their mutual links, discovered that their behaviour is not random, and they identified few web pages which were highly connected, whereas the majority appeared to be quite sparse.

In particular, they are characterised by a degree distribution which follows a power law. Formally speaking, this can be formalised, as

$$pk \approx k\text{-}\gamma$$

where pk is the fraction of nodes in the network with degree k, and γ is a parameter which has been empirically shown to be usually in the range $2<\gamma<3$ (Albert, et al. 2002).. In scale-free networks there is the relatively high likelihood to have *hubs*, which govern the behaviour of the information flow and how it propagates through the nodes (Albert, et al. 2002)..

Another important aspect of scale-free is that they are considered as *evolving*. In other words, they can successfully model systems that exhibit dynamical properties. This includes scenarios where nodes and edges are created over time, following the *preferential attachment* property (Albert, et al. 2002)., which states that new nodes are more likely to connect to existing nodes with higher degrees. A rather simple, yet explanatory example, is the fact that highly connected individuals tend to have more connections over time. In this paper, we do not investigate the dynamics of topologically reduced networks, since the data-set we have investigated consists of "static" entries. However, one of the main challenges in Big Data is the ability to process information that changes within specific time constraints. This will be the focus of future research.

In the next section, we discuss the main components of our approach based on the topological structures described above.

Big Data

Data is continuously generated in different formats and sizes, which introduces critical challenges as they need to be addressed utilising appropriate algorithms ant technology.

The main properties that characterise Big Data include:

- **Volume.** There is a huge quantity of data that is accessible, consisting of real-time and historical data.
- **Velocity.** Real-time data is continuously created and changed, which requires suitable techniques to process and assess them within specific time constraints.
- **Variety.** Data is created in all shapes and forms, depending on their source such as structured or unstructured.
- **Veracity.** Data tend to contain contradictory and missing information which needs to be addressed in order to provide relevant intelligence.

The ability to address most (if not all) of these components, is at the very core of Big Data research. Furthermore, due to the ubiquity of data, applications of Big Data have increasingly drawn attention from the academic and business world. In fact, its potential has proved to be huge with relevant and important applications.

Description of our Approach

As discussed above, the assessment and management of Big Data provide unprecedented challenges as well as opportunities. The aim of this paper is to introduce a method to identify the relational structure which is created by data, so that it can be implemented accurately and in a computationally efficient manner. More specifically, the elements of the data sets and their mutual relationships define real-world networks. Clearly, such networks can vary in size, from relatively small and highly connected structures, to huge networks with several thousands of nodes and edges, which can also incorporate erroneous information.

The ability to topologically reduce real-world networks addresses the above challenges, by identifying and ranking the best network structure that can approximate data structure (Watts, at al., 1998).

Broadly speaking, the main components include the following steps:

- Definition of a network based on the relations among elements of a dataset. These are identified either directly from structured entries, or via text mining techniques.
- Identification of two different *semi-isomorphic* networks (i.e. networks with same nodes), namely scale-free, and random.
- The main parameters characterising the above networks are then identified to attempt to approximate the topology of the original network.
- Finally, the above approximations are ranked according to their accuracy.

Figure 1 depicts the different stages of our method as described above. Note that the main motivation of this approach is to determine whether a real-world network exhibits a purely random network, implying a randomly distributed co-existence of the nodes, or a scale-free structure. In particular, this would suggest a pattern in the node co-occurrence.

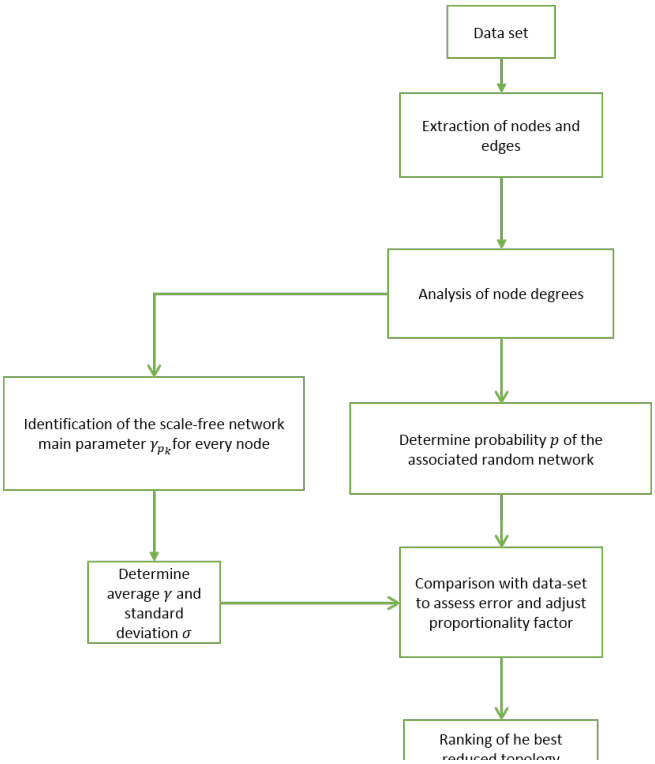

Figure 1: Scale-free network algorithm

Generation of Networks

As already mentioned above, the topological reduction of real-world networks has several benefits. In fact, extracting the structure of potentially big data-sets allows a better understanding and manipulation of the information embedded in such data-sets. However, due to the diverse and complex nature of Big Data, there is a need to balance accuracy with computational efficiency.

In Section Theoretical Background, the general properties of random and scale-free networks were introduced. In order to topologically reduce a real-world network, it is crucial to determine the parameters that govern the corresponding structure.

The behaviour of the node degrees in scale-free networks is based on the parameter γ, which can be easily found for the nodes with the degree, namely

$$\gamma pk = -\log pk \log k$$

where the suffix refers to the proportion pk of nodes with degrees k. If we consider all the different subsets of nodes defined by the node degree, we then have a sequence γpk1, …, γpkj , whose average is

$$\Delta = l = 1j \gamma pkln$$

where n = |V|, is the cardinality of the node set. Its standard deviation is then

$$\sigma = 1nl = 1j(\gamma pkl - \Delta)2$$

The average Δ and the standard deviation σ are used as the main parameter of our algorithm to estimate the value of γ, as well in the ranking process as discussed in Section Ranking of the Generated Networks.

More specifically, the corresponding algorithm is as follows

Scale-free Network Algorithm (m,n)

```
temp = 0
for i = 1 to m
      do
                  γpk  = -logp(k)logk
```

```
                temp = γ+ temp
Δ= temp/n

for each γ
       do
                find σ

return Δ, σ
```

where m is the number of subsets of V with the same degree, i.e. pki for $i=1,\ldots,m$, and n is the cardinality of V.

In a similar fashion, we can algebraically manipulate the equation that governs the degree distribution in random network, namely

$$pkk!(n-1)k=pk\ e1-np$$

to obtain a more computationally efficient approach.

As discussed in (Trovati, et al., 2014), we can re-write the above equation as

$$logpkk!(n-1)k=k\ logp+\ 1-np$$

which can be approximated as

$$2p2-1-nk+4+\ log2pk\ k!n-1k+0.5,$$

by expanding $logp$ at $p = 0.5$, as

$$logp\approx 2p-\ log2-2(p-0.5)2$$

Similarly to the above, we have developed the following algorithm

Random Network Algorithm (m,n)

```
       temp = 0
       for i from 1 to m
              do
                                                                    solve(
2p2-1-nk+4+
log2pk k!n-1k+0.5)
                     if p >= 0 and p <= 1
                          do
                                 temp = temp + p
Δ= temp/n
       for each γ
              do
                     find σ

       return Δ, σ
```

An important constraint is that p has to be within the interval $[0, 1]$ for obvious reasons. Any value of p outside the above interval will be ignored.

Text and Sentiment Analysis of Data Sets

Big Data come in a multitude of types, and can be structured as well as unstructured. As a consequence, we also consider specific textual extraction capabilities to address a wider range of data-sets. In particular, extracting the relevant information from textual data enables to successfully isolate the nodes of the network that are associated with them. However, when data include textual information, the identification and analysis of their relations can be problematic due to the ambiguous nature of human language. Depending on the general context and information on the type of data and their structure, a variety of text mining techniques can be used.

Text Mining (De Marneffe, et al., 2006) is typically a computationally demanding task which is clearly not suitable when dealing with Big Data. There are a variety of techniques to extract information from text, including grammar and statistical based approaches. Sentiment analysis (Liu, 2012). , is a very active research field which aims to detect "moods", or "opinions", expressed by blogs, websites, texts, feeds, etc. to understand the type of emotions that are captured by textual information.

The dataset which we have used for validation, as discussed in Section Evaluation, contains information on air accidents and near misses. In particular, some of the entries consisted of pilots' comments which would be particularly suitable for the above method.

In this paper, we apply a method based on sentiment analysis defined by a vocabulary consisting of specific keywords. These were semi-automatically determined by creating a large set of keywords and cue phrases by extracting them from the tagged version of the Brown Corpus, containing approximately 500 samples of English-language texts (Trovati, et al., 2015) .

Table 1: A small selection of keywords as described above

Abandon	Calamity
Abnormal	Collision
Abrupt	Delay
Accident	Disaster
Quit	Risky
Turbulence	Warning
Caution	Cancel
Challenge	Complain
Complicate	Confusion
Damage	Danger
Defective	Deteriorate

Subsequently, the next steps include

1. Text fragments are first shallow parsed via the Stanford Parser (De Marneffe, et al., 2006), to allow a computationally feasible syntactic analysis.
2. A grammar-based extraction extracts triples of the form `(NP, verb, keyword)`, where
    - `NP`, or noun phrase, contains the subject of the sentence. Only the head noun is isolated,
    - `verb` is the linking verb, and
    - `keyword` consists of one or more keywords as mentioned above.

All the identified triples are used to populate the nodes and edges of a network, which capture the corresponding relationships among the different data entries.

Ranking of the Generated Networks

As mentioned above, the aim of the ranking is to suggest which network, or networks, are the most appropriate to approximate the given dataset.

The above algorithms provide an approximated value of the most relevant parameters, namely γ and p for a scale-free and a random network respectively, as well as the associated standard deviations.

In this paper we do not provide a fully automated ranking system, as we require human intervention for this particular task. More specifically, the user has to assess which network is most suitable according to the features of the parameters, and their corresponding standard deviations. For example, if we obtain a value with a large standard deviation, we might assess it as not an accurate topology reduction.

Another important aspect if this method is the analysis of long tails distributions for scale-free networks, which tend to exhibit such property due to their exponential nature. However, experimental evaluations indicated that long tails might lead to inaccurate γ evaluation (Trovati, et al., 2014).

As a consequence, we also included the following algorithm which returns the interval in which we have values that do not yet follow a log-tail distribution.

Long-Tail Algorithm (value(x),d)

```
n = initial_value
while abs(value(n+1) – value(n))<d
      n = n+1
return n
```

where value(x) corresponds to the number of nodes with n degree.

Description of the Dataset

The dataset we have analysed consists of large data-set containing earthquake occurrences, which have been collected by the European-Mediterranean Seismological Centre (European-Mediterranean Seismological Centre Database, 2014). More specifically, it includes real time parametric provided by 65 seismological networks of the Euro-Med region, recorded from October 2004 to September 2009, with a magnitude up to 9.3. The collected data are subsequently stored in a large database, which contains structured data.

In (Trovati, et al., 2014), the authors consider a data-set about air accidents and near misses, available from (Aviation Safety Reporting System Database, 2014), which contains unstructured entries in the form of textual information. This was analysed by using the method discussed in above. Even though we did not carry out a formal evaluation of the text extraction results, we noticed that it produced an accuracy above 70%. In future research, we are aiming to expand the text mining extraction and carry out a comprehensive evaluation across a variety of data-sets.

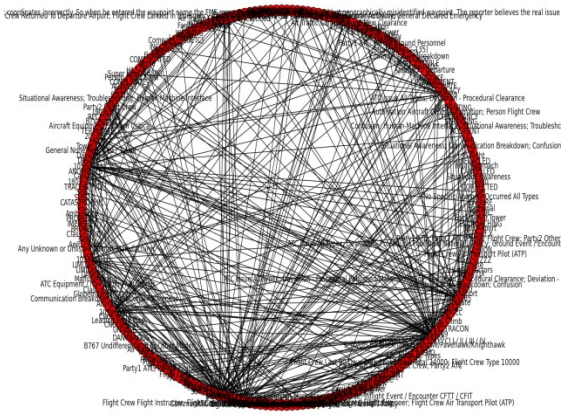

Figure 2: Scale-free network topologically reduced from (Akoumianakis, et al., 2012).

Table 2: Example of the nodes extracted from specific textual entries, as described in (Akoumianakis, et al., 2012).

Entry in data-set	Sentiment analysis output
ZID Controller issued a 'direct' to destination clearance when in fact the destination was unclear resulting in the aircraft turning to an unexpected heading.	**Node 1** : destination unclear **Node 2**: unexpected heading
A pilot reported his aircraft's FMS recalled an Engine Out	**Node**: incorrect geographical coordinates

waypoint with incorrect geographical coordinates because a previous crew had manually entered a correctly named waypoint with incorrect coordinates.	
ONT Controller described multiple TCAS RA events; resulting in two go arounds during police helicopter operations near the landing runway.	**Node 1**: multiple TCAS RA events **Node 2**: two go arounds

Evaluation

In order to carry out an evaluation of our approach, the data were suitably manipulated to produce a network, or in other words, nodes were associated to the different entries whose mutual relational connections defined the edges of the network.

In particular,

1. The textual entries were analysed as described in Section Text and Sentiment Analysis of Data Sets to extract triples of the form `(NP, verb, keyword)`. These identified nodes as entities captured by the `NP` linked to specific keywords associated to a state. Note that it may contain a fragment of a sentence. As a consequence, we only extract the head noun, including the corresponding adjectives and quantifiers. This defined nodes, which were also connected to the other entries in the data set according to its hierarchical structure.
2. This generated a (non-fully connected) network with 47,593 nodes and 65,536 edges, with an average degree of 2.75.
3. We then considered specific parameters, namely the date of the earthquake activity, its geographical location, time of the day, and its intensity, and assessed their corresponding reduced-topology networks.
4. The algorithms described in Section Generation of Networks were implemented to assess whether a scale-free or random topology would best approximate the network, if any at all. Note that this evaluation, as mentioned above, is not fully automated, and it ultimately depends on the user's judgement, supported by the values of the different parameters, and corresponding standard deviation.

Table 2 shows all the relevant parameters associated with the corresponding networks produced by the algorithms described earlier. It can be seen that the best values, considering their standard deviations, appear to indicate that the original network exhibits scale-free properties. This further supports recent research which suggests that no real-world network is likely to be purely random (Albert, et al. 2002).

Table 2: Evaluation of results for the three different parameters.

Parameter 1: Date of seismic activity	**Scale-free Network**	γ= 2.17 σ= 1.63
	Random Network	p=0.17 σ=2.2
Parameter 2: Geographical location of seismic activity	**Scale-free Network**	γ= 2.74 σ= 0.73
	Random Network	p=0.09 σ=0.81
Parameter 3: Time of seismic activity	**Scale-free Network**	γ= 2.93 σ= 1.98
	Random Network	p=0.21 σ=4.9
Parameter 4: Intensity of seismic activity	**Scale-free Network**	γ= 2.89 σ= 1.96
	Random Network	p=0.13 σ=5.3

In fact, according to the ranking system proposed in (Trovati, et al., 2014), scale-free is the network structure with reduced topology that provides the most accurate approximation of the original network across the parameters considered. For example Figure 2, clearly showsthe exponential nature of the node degrees with respect to geographical location of the seismic activity..

In particular, figure 6 depicts how the Long-Tail algorithm produces more accurate results with respect to the pre-long tail data, proposed in the existing literature (Clauset, et al., 2009).

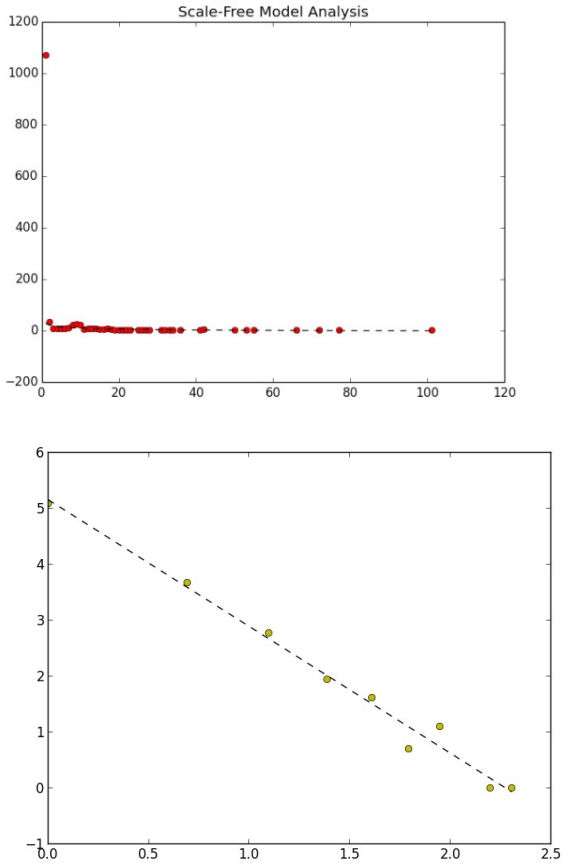

Figure 2: Plots of the degrees of nodes against number of edges based on the geographical location of seismic activity. In particular, the latter is a log scale, which shows the exponential nature of the degree distribution.

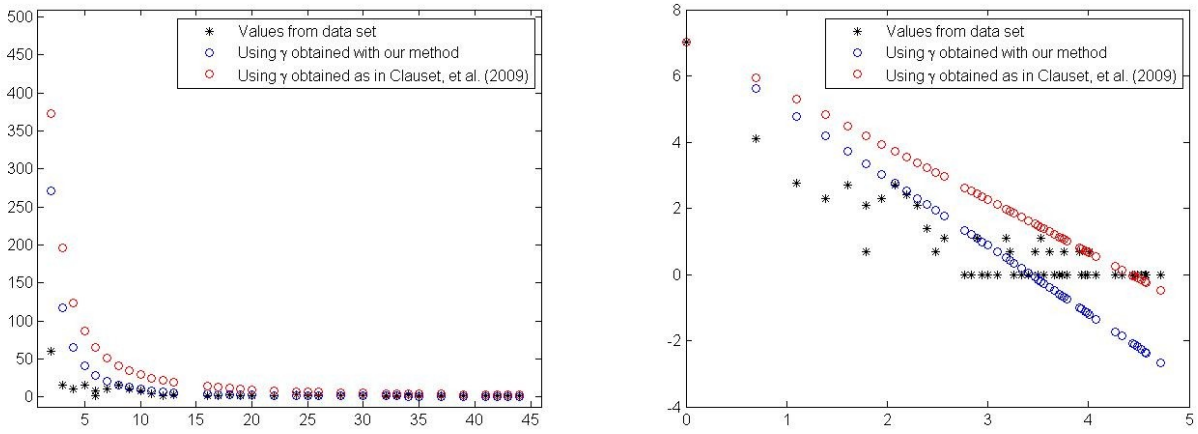

Figure 6: Comparison of the results when using the Long-Tail algorithm (Clauset, 2009) in the case of nodes against number of edges based on the date of seismic activity. The latter plot is in log scale, which again, clearly shows the exponential nature of the corresponding network.

The values of γ across the four different networks vary between 2.17 and 2.94 with cut-offs at an approximate degree $n = 20$. As discussed in (Newman, 2003), there is strong and increasing evidence that real-world networks exhibit scale-free topology with $2<\gamma<3$. Our evaluation further reinforce this, showing that seismic activity (European-Mediterranean Seismological Centre Database, 2014) appears to follow such trend.

Even though we did not directly evaluate the computational efficiency of our algorithms, we noted that the overall performance showed evidence of effectiveness. A full evaluation of the performance of our approach will be carried out in future research.

Conclusions and Future Steps

In this paper, we have discussed the evaluation of the method we introduced in (Trovati, et al., 2014) on a real-world big dataset as described in Section Description of the Dataset. Our results show that the corresponding network can be topologically reduced to a scale-free network according to the parameters discussed in Section Evaluation, with $2<\gamma<3$. This further supports the fact that real-world networks exhibit this particular structure (Albert, et al. 2002).

We are planning to expand and enhance our algorithms to address more topological structures, and different dataset types. In particular, we are planning to expand our method to address the challenges posed by huge datasets including data inconsistencies, as well as missing information to define full networks. We will also directly address computational efficiency of our approach to ensure its full scalability.

References

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

Albert R, Jeong H, and Barabasi, A. (1999). Diameter of the World Wide Web. In: Nature 401, pp. 130–131.

Bollobas B. (1998). Modern Graph Theory. Graduate Texts in Mathematics, Vol. 184, Springer, New York.

P. Erdȍs P and A. Rényi A. (1960). On the evolution of random graphs. Magyar Tud. Akad. Mat. Kutaté Int. Közl., 5:17–61.

Humphries M D, Gurney K. (2008). Network "Small-World-Ness": A Quantitative Method for Determining Canonical Network Equivalence. PLoS ONE, vol 3:4

Newman M. The Structure and Function of Complex Networks. (2003). SIAM Review, vol. 45:2, pp. 167-256.

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

Akoumianakis D, Karadimitriou N, Vlachakis G, Milolidakis G and Bessis N. (2012). Internet of Things as Virtual Settlements: Insights from Excavating Social Media Sites. INCoS 2012: 132-139

Trovati M, Bessis N, Huber A, Zelenkauskaite A, Asimakopoulou E. (2014). Extraction, Identification and Ranking of Network Structures from Data Sets. Proceedings of CISIS.

Zelenkauskaite A, Bessis N, Sotiriadis S and Asimakopoulou E. (2012). Interconnectedness of Complex Systems of Internet of Things through Social Network Analysis for Disaster Management. Proceedings of INCoS 2012: 503-508

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

Milgram S. (1967). The Small World Problem, Psychology Today, Vol. 2, 60–67.
Aviation Safety Reporting System Database. (2014). Available from http://asrs.arc.nasa.gov/search/database.html. [1 March 2014]
De Marneffe M F, MacCartney B and Manning C D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses, LREC 2006.
Liu B. (2012). Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers.
Trovati M, Asimakopoulou E, Bessis N. (2014). An Analytical Tool to Map Big Data to Networks with Reduced Topologies, Proceedings of InCoS 2014.
Trovati M, Bessis N. (2015) An Influence Assessment Method Based on Co-Occurrence For Topologically Reduced Big Data Sets, Submitted to Soft Computing.
Dingli A., and Seychell D. (2012). Taking Social Networks to the Next Level, IJDST, Vol. 3.
Clauset A., Shalizi C.R., and Newman M.E.J. (2009). Power-law distributions in empirical data, SIAM Review 51(4), 661-703.