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

Marcello Trovati, Department of Computing and Mathematics, University of Derby, Derby, 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, the author 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, the author will use a large dataset containing information on the seismologic activity recorded by the European-Mediterranean Seismological Centre. The author will show that it provides an accurate, agile, and scalable tool to extract useful information. This further motivates his 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: *Data Analytics, Information Extraction, Knowledge Discovery, Networks, Seismological Data, Text Mining*

1. 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

DOI: 10.4018/IJDST.2015040102

Copyright © 2015, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

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* $V = \{v_i\}_{i=1}^{n}$, which are connected as specified by the *edge set* $E = \{e_{ij}\}_{i \neq j=1}^{n}$ (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, 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 II we describe the main features of the networks considered in the paper. Section III discusses the relevance of Big Data and its properties, while Section IV focuses on the description and implementation of the main algorithms. Finally, Section V discusses the evaluation we have carried out, and Section VI concludes and prompts future directions of our work.

Copyright © 2015, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

13 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/article/reduced-topologically-real-world-networks/126175

Related Content

A Theoretic Representation of the Effects of Targeted Failures in HPC Systems

A. Don Clark (2016). *Innovative Research and Applications in Next-Generation High Performance Computing (pp. 253-276).*
www.irma-international.org/chapter/a-theoretic-representation-of-the-effects-of-targeted-failures-in-hpc-systems/159048

Evaluation of Encryption Procedure for User Attestation System Using a Cellular Phone

Noriyasu Yamamotoand Toshihiko Wakahara (2012). *International Journal of Distributed Systems and Technologies (pp. 15-26).*
www.irma-international.org/article/evaluation-encryption-procedure-user-attestation/67555

Data Mining in Proteomics Using Grid Computing

Fotis Psomopoulosand Pericles Mitkas (2012). *Grid and Cloud Computing: Concepts, Methodologies, Tools and Applications (pp. 918-940).*
www.irma-international.org/chapter/data-mining-proteomics-using-grid/64522

EcoGrid: A Toolkit for Modelling and Simulation of Grid Computing Environment for Evaluation of Resource Management Algorithms

Hemant Kumar Mehta (2014). *International Journal of Grid and High Performance Computing (pp. 1-16).*
www.irma-international.org/article/ecogrid/119449

Watermarking of EEG Data to Provide Security Based on DWT-SVD and Optimized by Firefly Algorithm

Akash Kumar Gupta, Chinmay Chakrabortyand Bharat Gupta (2022). *International Journal of Distributed Systems and Technologies (pp. 1-16).*
www.irma-international.org/article/watermarking-of-eeg-data-to-provide-security-based-on-dwt-svd-and-optimized-by-firefly-algorithm/307902