Computer Science Review 29 (2018) 1–13

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Review article

Change point detection in social networks—Critical review with experiments

Lucy Kendrick [a], Katarzyna Musial [b,*], Bogdan Gabrys [b]

[a] Data Science Institute, Faculty of Science and Technology, Bournemouth University, BH12 5BB, Bournemouth, UK
[b] Advanced Analytics Institute, Faculty of Engineering and IT, University of Technology Sydney, Australia

ARTICLE INFO

Article history:
Received 13 December 2017
Received in revised form 2 May 2018
Accepted 2 May 2018

Keywords:
Change point detection
Social network
Dynamic network

ABSTRACT

Change point detection in social networks is an important element in developing the understanding of dynamic systems. This complex and growing area of research has no clear guidelines on what methods to use or in which circumstances. This paper critically discusses several possible network metrics to be used for a change point detection problem and conducts an experimental, comparative analysis using the Enron and MIT networks. Bayesian change point detection analysis is conducted on different global graph metrics (Size, Density, Average Clustering Coefficient, Average Shortest Path) as well as metrics derived from the Hierarchical and Block models (Entropy, Edge Probability, No. of Communities, Hierarchy Level Membership). The results produced the posterior probability of a change point at weekly time intervals that were analysed against ground truth change points using precision and recall measures. Results suggest that computationally heavy generative models offer only slightly better results compared to some of the global graph metrics. The simplest metrics used in the experiments, i.e. nodes and links numbers, are the recommended choice for detecting overall structural changes.

Contents

* Corresponding author.
E-mail addresses: i7949850@bournemouth.ac.uk (L. Kendrick), katarzyna.musial-gabrys@uts.edu.au (K. Musial), bogdan.gabrys@uts.edu.au (B. Gabrys).

https://doi.org/10.1016/j.cosrev.2018.05.001
1574-0137/

1. Introduction

For many years the analysis of complex networks remained a static exercise. Now research is increasingly viewing networks as dynamic systems, where the dynamic properties are as important as overall network structure. The computational capability to study not only large graphs, but a long sequence of large graphs over time has led to growing research in the field of detecting, modelling and predicting changes in complex networks [1–7]. The focus of this paper is on the problem of change point detection, which is a form of dynamic anomaly detection that has a long history of study in traditional time series datasets [8–13].

There are many detection algorithms to find individual anomalies in static graphs [2]. These focus on the more traditional form of an anomaly that involves finding one unusual data point or node. The motivation behind this paper stems from the growing field of research that uses generative models to study change point detection in dynamic networks [14,15,3,4,6,1]. Generative models are ways to probabilistically represent network data into sets of communities or hierarchy. It offers a potentially rich representation that can monitor smaller or subtle changes happening in subsections of a graph.

As a new area of research there is a need to establish the best ways to model the change point detection problem. There is also a lack of understanding in the generative model space of why one type of model should be selected over another. The aim of our research has therefore been to critically review the existing approaches and conduct an experimental analysis exploring different potential network metrics that can be used to detect changes in such complex, dynamic networks.

The paper begins with a review of the related work in Section 2 that provides a discussion on change point detection and the use of generative models in this research area. This is followed by Section 3 describing the metrics used in the experimental analysis. The datasets, experimental set up, the results of conducted experiments and the related discussions are presented in Section 4. Finally, Section 5 provides the conclusions and highlights some identified future research directions.

2. Related work

The problem of Change Point Detection (CPD) historically stems from research assessing classical time series data to identify a change in the underlying mean or distribution of a given variable. Changes can be identified from calculations that measure the posterior probability of a change in monitored parameters. Such techniques have been successfully applied to many engineering and control problems to identify faults in systems [8,13]. The overriding aim for CPD research, in the field of complex networks, is to identify a point in time where the graph exhibits a difference in behaviour. This time point can then be analysed to uncover an underlying cause.

Change Point Detection in complex networks is often tied to the field of anomaly detection. Both research areas use similar methods that exploit the existence of communities in graphs to establish unusual behaviour [2]. As a relatively new area of research there is no leading methodology used to conduct CPD in networks. According to a common methodology for CPD using time series analysis, the first step should be a preliminary investigation of the best way to model the problem followed by a selection of the best variables to be used as change indicators [8].

From the literature we find that change point and anomaly detection research will often use generative network models as a way to model the problem on a complex network. Generative models provide a well-recognised way of finding community structures or hierarchy in a graph with the additional benefit of using probabilistic values. Though most CPD studies agree on the use of generative models in this research area, they do not agree on any specific one to be clearly better than the others.

2.1. The change point detection problem

In the context of statistical methods employed, Basseville et al. [8] define three main problem areas in CPD:

– On-line-detection, where it is required that the change be identified as soon as possible to near real time. In the context of control problems this is often the main aim. This would ensure any faults in a monitored system caused by an unforeseen change can be highlighted instantly. This method, however, suffers from the issue of false alarms (false positives) where what may appear to be a change was only an anomaly.

– Off-line hypothesis testing, where the aim is to maximise the trade off between correctly identified change points and false alarms. This is often used as a retrospective analysis. This method has been often used as evidenced in the literature reviewed in the following sections.

– Detecting the exact time of a change, which can be used in combination with the above two approaches but where only one change point is to be discovered and it is assumed that no other change has taken place within the analysed section of data. This would be very important to a more time-sensitive application (on-line analysis) or where the real time detection is not important (off-line detection) but the exact moment of change is needed for further analysis.

2.2. Change point detection methods in time series data

There are well developed methodologies for finding change points in traditional time series data, where a metric is monitored over a number of time bins and evaluated for change. There is a number of methods utilising different data mining techniques which broadly search for abrupt change in the mean or variance of the monitored variables/data. One of such methods, which is used in our experiments, is a Bayesian Change Point (BCP) detection that works under the assumption that the underlying sequence of time series data can be partitioned into a sequence of blocks. Within each of these blocks the data exhibits behaviour described by a set of parameters whose values do not change between blocks. BCP techniques often cite the use of product partition models which are defined based on the assumption that observations within each random partition have independent prior distributions [10]. The number of blocks in the data is unknown and is randomly sampled using the Monte-Carlo technique [9]. The main metric to determine the change event is the posterior probability of change that is equated to an increasing change in a given parameter between the defined bins [11]. [12] is a popular, more recent study that tackles the change point problem from an on-line perspective with time series datasets. The work is based on the previously mentioned assumption that the sequence can be divided into partitions