Physica A 448 (2016) 235–247

Contents lists available at ScienceDirect

Physica A

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

What is the dimension of citation space?

James R. Clough *, Tim S. Evans

Centre for Complexity Science, Imperial College London, South Kensington campus, London, SW7 2AZ, UK

HIGHLIGHTS

- Causal constraints in network structure require the development of novel methods.
- Tools from causal set theory can be applied to study citation networks.
- This can reveal differences in citation behaviour between subfields.
- We measure the dimension of the spacetime each citation network is closest to.
- Other directed acyclic graphs can be studied with these methods.

ARTICLE INFO

Article history:
Received 24 April 2015
Received in revised form 24 September 2015
Available online 23 December 2015

Keywords:
Citation network
Directed acyclic graph
Minkowski space
Dimension
Causal set
Network geometry

ABSTRACT

Citation networks represent the flow of information between agents. They are constrained in time and so form directed acyclic graphs which have a causal structure. Here we provide novel quantitative methods to characterise that structure by adapting methods used in the causal set approach to quantum gravity by considering the networks to be embedded in a Minkowski spacetime and measuring its dimension using Myrheim–Meyer and Midpoint-scaling estimates. We illustrate these methods on citation networks from the arXiv, supreme court judgements from the USA, and patents and find that otherwise similar citation networks have measurably different dimensions. We suggest that these differences can be interpreted in terms of the level of diversity or narrowness in citation behaviour.

1. Introduction

Citation analysis has great potential to help researchers find useful academic papers [1], for inventors to find interesting patents [2], or for judges to discover relevant past judgements [3]. It is not, however, enough to simply count citations, because they can be made for a variety of reasons beyond an author genuinely finding a document useful [4–6]. To interpret the information encoded in a citation network we must be able to identify and describe the important features, such as the fat-tailed citation distributions [7–12], clustering [13], motifs [14], distance measures [15] and others [16,17]. Beginning with Price's cumulative advantage principle [18,19] there have been numerous attempts to construct models which replicate some of these features [9,20–26] and in doing so highlight potentially important mechanisms for the flow of information which drives the growth of real citation networks.

Citation networks are constrained in time, because authors can only cite something that has already been written.[1] This causal constraint prevents closed loops of directed edges in the graph, since all edges must point the same direction in time,

* Corresponding author.
E-mail address: jrc309@ic.ac.uk (J.R. Clough).

[1] Occasionally this is not the case for real citation networks. For instance, two authors may share and cite each others work before either is published, leading to two papers which both cite each other, clearly forming a cycle. Such 'acausal' edges are rare, making up less than 1% of edges in all citation networks considered here, and so were removed from the network since many techniques used here assume that the network forms a DAG.

http://dx.doi.org/10.1016/j.physa.2015.12.053
0378-4371/

and is the same constraint placed on causally connected events in physics. Therefore they naturally form Directed Acyclic Graphs (DAG) where a directed edge going from node A to node B represents document A having cited the document B.

By taking constraints into account it is often possible to create new methods of characterising network structure as is well known for networks embedded in space [27–29]. This kind of geometric approach can also be used by postulating the existence of a hidden geometric space which describes some aspects of a network's structure [30–32]. It is our view that the same is true of citation networks (as well as other networks which naturally form directed acyclic graphs because their edges represent causal connections). The role of time in citation networks means that instead of an underlying Riemannian manifold (such as usual Euclidean space of spatial networks, or a Hyperbolic space of Ref. [30]) we will consider a Lorentzian manifold.

In a Lorentzian manifold one dimension, usually representing time is treated differently to the other, spatial dimensions. Relations between points can be classified as timelike, null, or spacelike. The simplest such manifold, and the one we will consider here is Minkowski space which is the geometry of Einstein's special relativity, in which timelike relations correspond to causal connections. We will consider applying this geometry to DAGs by equating the directed edges in the graph with timelike separation between the nodes. With this approach we will characterise network structure using tools from the causal set approach to quantum gravity, in which spacetime (which is a Lorentzian manifold) is discretised and has the structure of a DAG. In particular, we will use methods which estimate the dimension of a Minkowski space from its causal structure and apply them to citation networks.

The rest of this paper is structured as follows. In Section 2 we introduce in detail the Lorentzian perspective of DAGs, seeing how they can be embedded in space and time, and the methods of estimating their dimension. In Section 3 we will adapt these methods for use on citation networks and test them on examples from academic papers, patents and court judgements. In Section 4 we interpret these results in terms of using dimension as a measure of citation diversity, and finally in Section 5 we discuss applications and the similarities and differences of this approach with others in the literature.

2. Dimension estimates for spacetime networks

In the causal set approach to quantum gravity, spacetime is seen as a set of discrete points with a partial order relation, called a causal set whose structure approximates the continuous space we perceive. Whether causal sets are a useful approach to understanding the physical universe is not the focus of this paper and so we direct the reader to Refs. [33–35] for more details.

We will consider only the simplest spacetimes, D dimensional Minkowski spacetimes of one time dimension and $D-1$ spatial dimensions.[2] To create a causal set which approximates the structure of Minkowski space, we begin by randomly and uniformly scattering points in Minkowski space by randomly assigning each point an associated time t and spatial co-ordinates x_i. Two points are causally connected if and only if they are timelike separated, meaning the differences in their co-ordinates satisfy:

$$(\Delta t)^2 > \sum_i (\Delta x_i)^2. \tag{1}$$

If this relationship is satisfied we then say that the point with the larger/smaller t coordinate is in the future/past lightcone of the other. In special relativity it is this relationship that defines whether two events in spacetime can causally affect one another. The direction of the edges is determined by the causal/temporal ordering as given by the ordering of the time coordinates, and provides a uniquely defined causal relationship. To translate this structure into the language of networks, we say each point is a node, and we add edges between nodes which are causally connected, i.e. their coordinates satisfy Eq. (1). We will use the convention that all edges point backwards in time. This process necessarily generates a DAG since all edges point the same direction in time.

An **interval** [A, B] in a DAG is the set of nodes which can be reached from A (are in its causal past) in one direction, and from B in the other direction (in its causal future) [36] as in Fig. 1. The dimension estimates used here are defined on an interval in a DAG.

To illustrate these estimates we will use a simple network model in which points are scattered uniformly at random in an interval of Minkowski space. We first create two extremal points with time co-ordinates of 0, and 1 respectively, with all spatial co-ordinates equal to 0. We then add more nodes to the network by assigning a random time co-ordinate between 0 and 1, and random spatial co-ordinates between −0.5 and 0.5 and allowing this node to be in the network if it has edges to the two extremal nodes and so lie within the interval such that $G_D(N)$ is a network created by this process with N nodes, which are described by D coordinates. We will refer to these networks as **spacetime networks** though they are also known as cone spaces in the mathematics literature [37].

The number of spatial dimensions will determine the structure of the graph this process creates. Extra spatial dimensions add further terms to the summation on the right hand side of Eq. (1) and make it less likely that two points are connected.

2 It is possible to define other similar networks, such as a cube-space [41], a spacetime network using a curved spacetime [31] or geometric space with other rules [26]. Minkowski spacetime is the simplest, being defined by just one parameter D, the measurement of which we will use to characterise the network's structure.