What is the dimension of citation space?

James R. Clough Tim .S. Evans

Imperial College London
Centre for Complexity Science

Mathematics of Networks 2014

Why should we care about citation analysis?

- There's just too many papers too read
- We need ways of deciding which papers are likely to be useful to our research
- Citation analysis can provide a mechanism for quantifying this

Can we just count citations?

The simplest method of measuring usefulness is counting citations.
But not all citations mean the same thing

- Cite a paper because it was genuinely useful
- Cite their own paper
- Cite their colleague/friend's paper
- Reviewer inserts citation to their paper
- Author copies from the bibliography of another paper
- Cite well known paper in the field even if it was not useful to this work

Can we just count citations?

The simplest method of measuring usefulness is counting citations.
But not all citations mean the same thing

- Cite a paper because it was genuinely useful
- Cite their own paper
- Cite their colleague/friend's paper
- Reviewer inserts citation to their paper
- Author copies from the bibliography of another paper
- Cite well known paper in the field even if it was not useful to this work

Can we just count citations?

The simplest method of measuring usefulness is counting citations.
But not all citations mean the same thing

- Cite a paper because it was genuinely useful
- Cite their own paper
- Cite their colleague/friend's paper
- Reviewer inserts citation to their paper
- Author copies from the bibliography of another paper
- Cite well known paper in the field even if it was not useful to this work

Can we just count citations?

The simplest method of measuring usefulness is counting citations.
But not all citations mean the same thing

- Cite a paper because it was genuinely useful
- Cite their own paper
- Cite their colleague/friend's paper
- Reviewer inserts citation to their paper
- Author copies from the bibliography of another paper
- Cite well known paper in the field even if it was not useful to this work

Can we just count citations?

The simplest method of measuring usefulness is counting citations.
But not all citations mean the same thing

- Cite a paper because it was genuinely useful
- Cite their own paper
- Cite their colleague/friend’s paper
- Reviewer inserts citation to their paper
- Author copies from the bibliography of another paper
- Cite well known paper in the field even if it was not useful to this work

Can we just count citations?

The simplest method of measuring usefulness is counting citations.
But not all citations mean the same thing

- Cite a paper because it was genuinely useful
- Cite their own paper
- Cite their colleague/friend’s paper
- Reviewer inserts citation to their paper
- Author copies from the bibliography of another paper
- Cite well known paper in the field even if it was not useful to this work

Can we just count citations?

The simplest method of measuring usefulness is counting citations.
But not all citations mean the same thing

- Cite a paper because it was genuinely useful
- Cite their own paper
- Cite their colleague/friend's paper
- Reviewer inserts citation to their paper
- Author copies from the bibliography of another paper
- Cite well known paper in the field even if it was not useful to this work

Is this a real problem?

- Academics and universities care about citation counts
- Journals care about impact factor
- Simkin & Roychowdhury estimated that around 80% of citations did not involve the author actually reading the paper they cite[2]

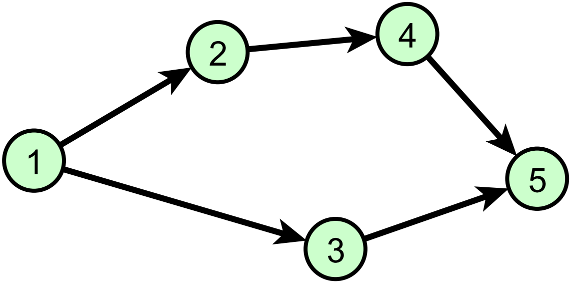
Solution? Use more of the network structure

- We have more information than just the number of citations a document has.
- There is a whole citation network structure to characterise and measure.
- Our approach - look at the causal structure of the network.

Citation Networks form Directed Acyclic Graphs

- We form a graph, where each document is a node
- A directed edge goes from node A to node B if A cites B in it's bibliography
- This means A must have been published *after* B, and edges go backwards in time. There can't be any closed loops (cycles).

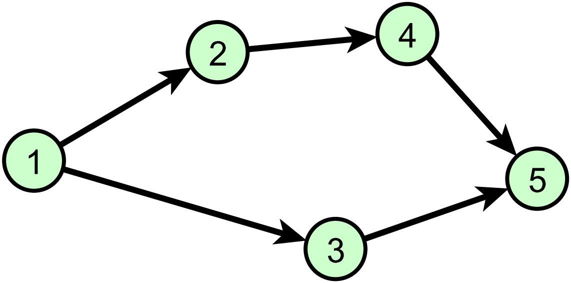
Citation Networks form directed Acyclic Graphs



Causal Structure

- Two nodes are causally connected if there is a path from one to the other, respecting edge direction.
- The set of these relations is what we mean by causal structure.

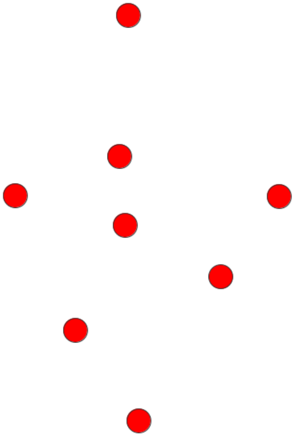
Causal Structure



Causal Structure

- We want to characterise the causal structure of a network
- We'll do this by making comparisons to the simplest set of models of networks with the same temporal constraints - networks embedded in Minkowski space.
- This model comes from a discrete approach to quantum gravity.

Take N nodes, and uniformly scatter them in a spacetime by giving them a time coordinate, t_α and $D-1$ spatial coordinates, x^i_α



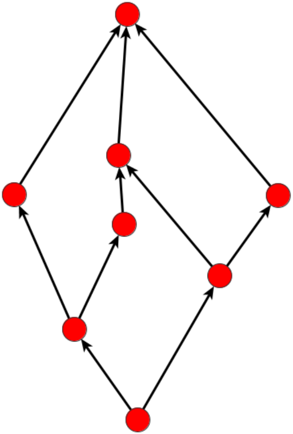
- Example of spacetime network where $D = 2$
- Time on vertical axis
- 1 spatial dimension on horizontal axis

- Take N nodes, and uniformly scatter them in a spacetime by giving them a time coordinate, t_α and $D-1$ spatial coordinates, x^i_α

- We then put an edge between two nodes, A and B if

$$(t_A - t_B)^2 > \sum_i (x^i_A - x^i_B)^2 \tag{1}$$

- So nodes are connected if they are more separated in time than they are in space
- Which is the same rule that defines how information can propagate through spacetime in special relativity.



- Example of spacetime network where $D = 2$
- Time on vertical axis
- $D - 1 = 1$ spatial dimension on horizontal axis
- Nearest neighbour links drawn for simplicity

Dimension

- Question - if we forget about the coordinates each point has, and just look at the nodes and edges can we work out what D was?

- Answer - yes - and this is how we will characterise these networks.

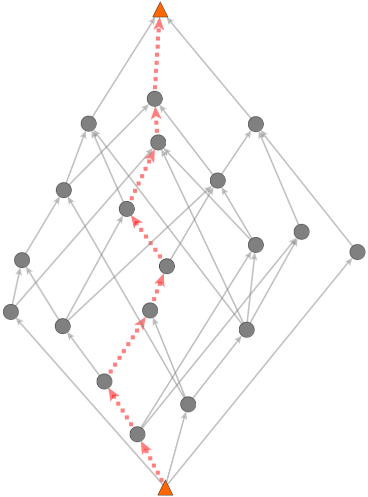
- Two ways of doing this developed in the causal set approach to quantum gravity. They only depend on the causal structure.

- While there are already methods of defining a ‘dimension’ for a network, they only consider spatial dimensions, but we will consider a time dimension separately as it has different constraints.

Method 1 - Midpoint Scaling Dimension

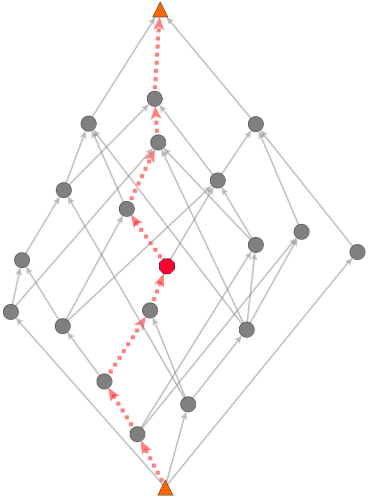
- Find a source node and a sink node - and look at the set of nodes in between them
- Find the longest chain from the source to the sink - this is a good approximation of the geodesic (shortest path) through that space

Method 1 - Midpoint Scaling Dimension



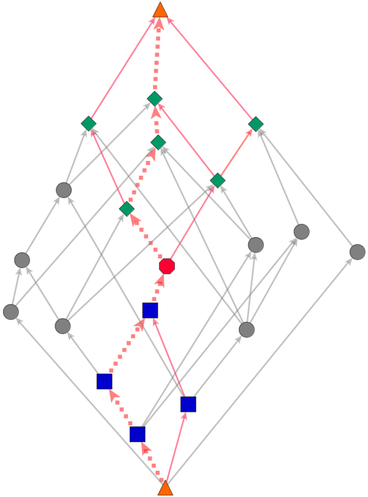
- Start and end nodes - orange triangles

Method 1 - Midpoint Scaling Dimension



- Start and end nodes - orange triangles
- Find midpoint - red octagon

Method 1 - Midpoint Scaling Dimension



- Start and end nodes - orange triangles
- Find midpoint - red octogon
- Find two intervals
- (start, middle) - blue squares
- (middle, end) - green diamonds
- The fraction of nodes in one of those intervals is $\frac{1}{2^D}$

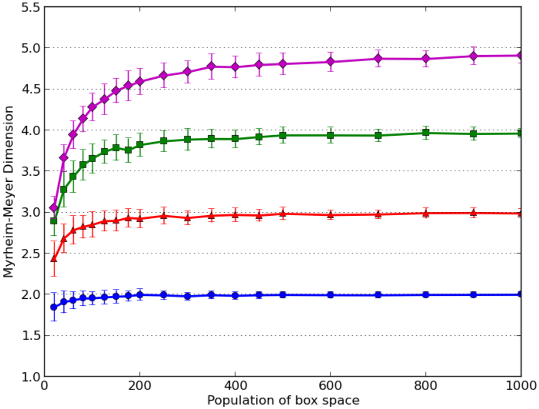
Method 2 - Myrheim-Meyer Dimension

Can be shown that the expected number of causally connected pairs, $\langle S_2 \rangle$ is just a function of N and D

$$\frac{\langle S_2 \rangle}{N^2} \equiv f(D) = \frac{\Gamma(D+1)\Gamma(D/2)}{4\Gamma(\frac{3}{2}D)} \tag{2}$$

So we can just measure how many of them there are, and then solve for D.[3]

OK - but does this actually work?



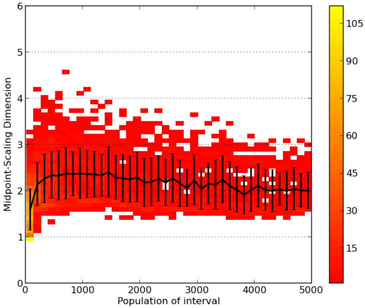
So what are we going to do?

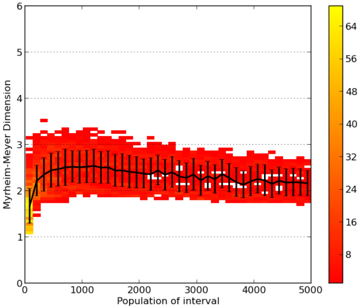
- So we have two different ways of measuring what kind of Minkowski space a network was embedded in
- We are now going to use these methods on real citation data and see what happens
- We will sample lots of intervals in the network to build up a picture of its causal structure

The data

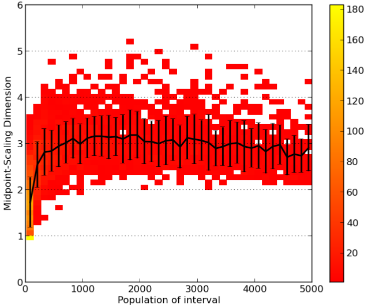
- Academic papers from the arXiv
- Patents from the USA (1970-2000)
- Judgements from the Supreme Court of the USA (1790-2012)

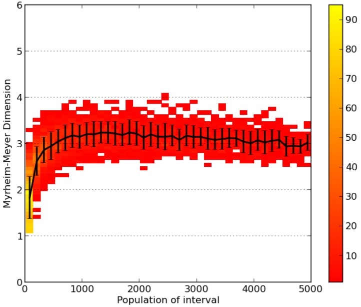
arXiv - high energy theory



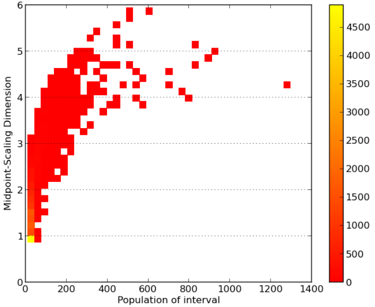


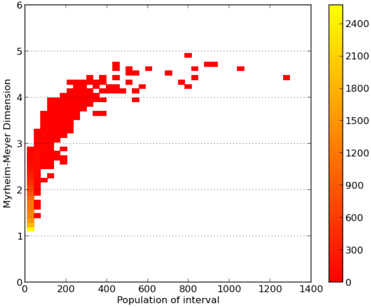
arXiv - high energy phenomenology



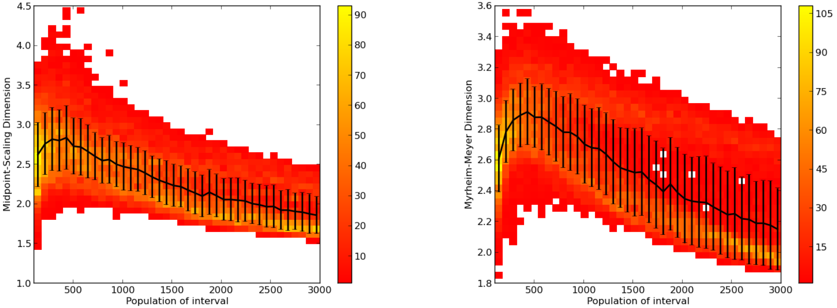


Patents


Midpoint-Scaling Dimension
0
1
2
3
4
5
6
0
200
400
600
800
1000
1200
1400
Population of interval
4500
4000
3500
3000
2500
2000
1500
1000
500
0


Myrheim-Meyer Dimension
0
1
2
3
4
5
6
0
200
400
600
800
1000
1200
1400
Population of interval
2400
2100
1800
1500
1200
900
600
300
0

Supreme Court



Interpretation

- High energy theory - 2
- High energy phenomenology - 3
- Patents - 5
- Supreme Court - 3 at small scales, 2 at large scales

Interpretation

- These dimension measures can easily distinguish between otherwise very similar citation networks
- They can be used to test whether models of citation network are really replicating the right behaviour on large scales

Interpretation

- We conjecture that these dimension measures can be interpreted in terms of how ‘broad’ or ‘narrow’ the citation behaviour in a field is.
- In a ‘narrow’ field where everybody cites all the same papers, the measured dimension would be 1
- In a ‘broad’ field where many people cite a paper without citing each other the measured dimension would be higher

Summary

- Independent dimension estimates agree, and there seems to be a consistently defined ‘spacetime dimension’ for citation networks
- They can be used to test whether models of citation network are really replicating the right behaviour, and can distinguish between similar networks
- Might help us ‘quantify interdisciplinarity’ - future work is on investigating this

Bibliography I

J.R. Clough, T.S. Evans
What is the dimension of citation space?
http://arxiv.org/abs/1408.1274
(Full list of citations available in this paper)

M.V. Simkin, V.P. Roychowdhury
Read before you cite!
arxiv.org/abs/condmat/0212043

D.D. Reid
Manifold dimension of a causal set
http://arxiv.org/abs/gr-qc/0207103