Understanding Main Path Analysis[†]

H. C. W. Price and T. S. Evans
Centre for Complexity Science & Abdus Salam Centre for Theoretical Physics, Imperial College London
27th February 2026

Keywords: Main Path Analysis, Directed Acyclic Graphs (DAGs), Citation Networks, Science of Science, Innovation Networks.

Motivation. Main Path Analysis (MPA) [1] is widely used to extract influential knowledge trajectories in citation, patent, and innovation networks represented as directed acyclic graphs (DAGs), see [2] for a recent summary. The use of Main Path Analysis is based on qualitative assessments: the assertion that the traversal counts, used to weight edges, capture flows of information, and post-hoc expert opinion that the results are reasonable. A principled foundation is needed to interpret knowledge–flow structures with confidence.

In our work [3] we ask if the MPA paths approximate meaningful "backbones" as represented by geodesics in some underlying geometry. We also look look for an information theoretical basis by constructing a new type of maximum entropy main path. Finally, we question the assumption that the collection of documents found on a single path is a robust way to find key documents, or if other structures would give better results.

Approach and Methodology. Our analysis includes analytic, geometric, and empirical aspects. By working on hypercubic lattices we derive results for traditional MPA paths and our new paths based on maximum entropy principles. This shows that both follow similar routes which are close to the geodesics.

We next study random geometric DAGs defined using Poisson point processes (see figure). By connecting nodes using an imposed order and Euclidean proximity we have a DAG embedded in a well defined geometry that also has some of the randomness associated with real networks.

Finally, we test our ideas on over twenty real citation networks including: arXiv sections, APS articles, US patents, US Supreme Court judgements, software dependencies, and vaccine innovation networks.

We compare paths with each other and, where available, to the geodesic. We also go beyond single paths and look at *baskets* of nodes, sets of nodes with the largest values of *criticality*. To do this we generalise the height $h^{(\mathrm{W})}(v)$ and depth $d^{(\mathrm{W})}(v)$ of each node v for a given edge weight W and from this we define *criticality* $c^{(\mathrm{W})}(v) = H^{(\mathrm{W})} - h^{(\mathrm{W})}(v) - d^{(\mathrm{W})}(v)$ where $H^{(\mathrm{W})}$ is the largest height of any node. The criticality is non-negative and is only zero for nodes lying on paths that caryy maximum weight.

Results. What we find is that both paths defined using traditional Main Path Analysis methods and those defined with our new entropy maximisation are similar to each other. We also find that these paths are very similar to the *longest path*, paths with the largest number of edges (equivalent to edge weight one).

However, we also show that these single paths miss large numbers of paths that are also optimal or which are close to optimal in terms of the criticality measures. This shows that traditional single-path methods are not robust, unlike our approach using 'baskets' of nodes.

[†]Extended abstract for APCNCS2026, to be presented by T. S. Evans, based on [3].

Conclusions and Outlook. We provide an information–theoretic and geometric foundation for Main Path Analysis and but we also show there is no practical difference from those paths found using simpler longest path methods. The reason is these paths all encode the fundamental geometry in the same way since they are all *maximising* some measure of path length. However, we also showed that single paths are not robust. So we conclude that using 'baskets' of nodes which lie on one of many longest paths (edge weight one) is a much more effective than Main Path analysis or our entropy-based approach to find documents which were critical to an innovation process in a citation network.

References

[1] NP Hummon, P Doreian. Connectivity in a citation network: The development of DNA theory. *Social Networks*, 11:39–63, 1989.

[2] JS Liu, LYY Lu, M Ho. A few notes on main path analysis. *Scientometrics*, 119:379–391, 2019.

[3] HCW Price, TS Evans. Understanding main path analysis. `arXiv:2512.12355`, 2025.

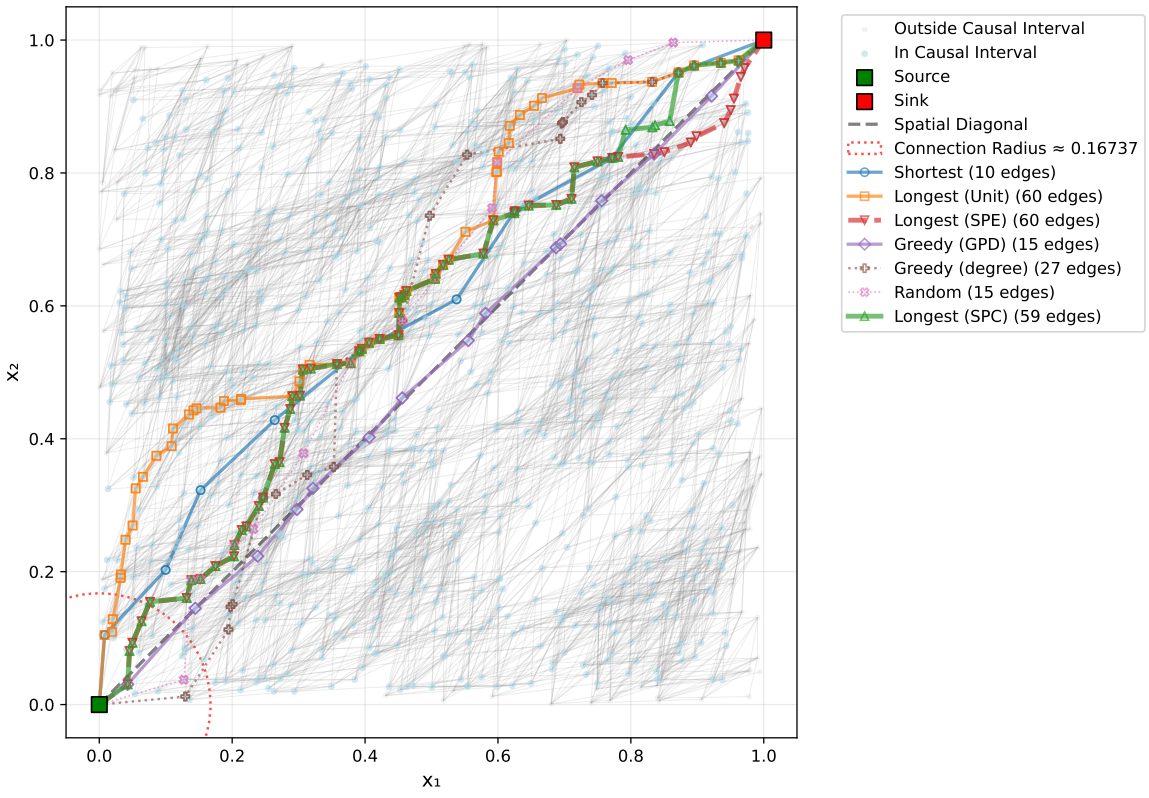

Figure 1: Random geometric DAG in two dimensions with $N = 1000$ box points and average degree 22.0 showing various path types from source (bottom left) to sink (top right) containing between 10 and 60 edges. By way of comparison, the diagonal is $\sqrt{2} \approx 44.8\,a \approx 8.47\,R$ where a is the typical inter-node spacing and R is the largest distance between points. SPC is the main path, SPE is our entropy version, whilst the longest path is by number of edges.