Clustered scale-free networks

Modelling the processes that create clustering and power-law distributions

Ewan Colman and Geoff Rodgers

Brunel University

Ewan.Colman@brunel.ac.uk

Abstract

Preferential attachment provides a simple explanation for the scale-free topology found in many networks. In many cases however, such as the network of citations in scientific literature, models are needed that also exhibit high levels of clustering. Here we discuss various growth processes for clustered complex networks. In particular, we consider a class of processes that are driven by *triad formation* by *link copying* and formulate this mechanism using parameters that can easily be inferred from citation data. Observing a reasonable amount of agreement between the results of the model and the data, we conclude that our research may be significance in quantifying the quality of scientific articles and other similar systems.

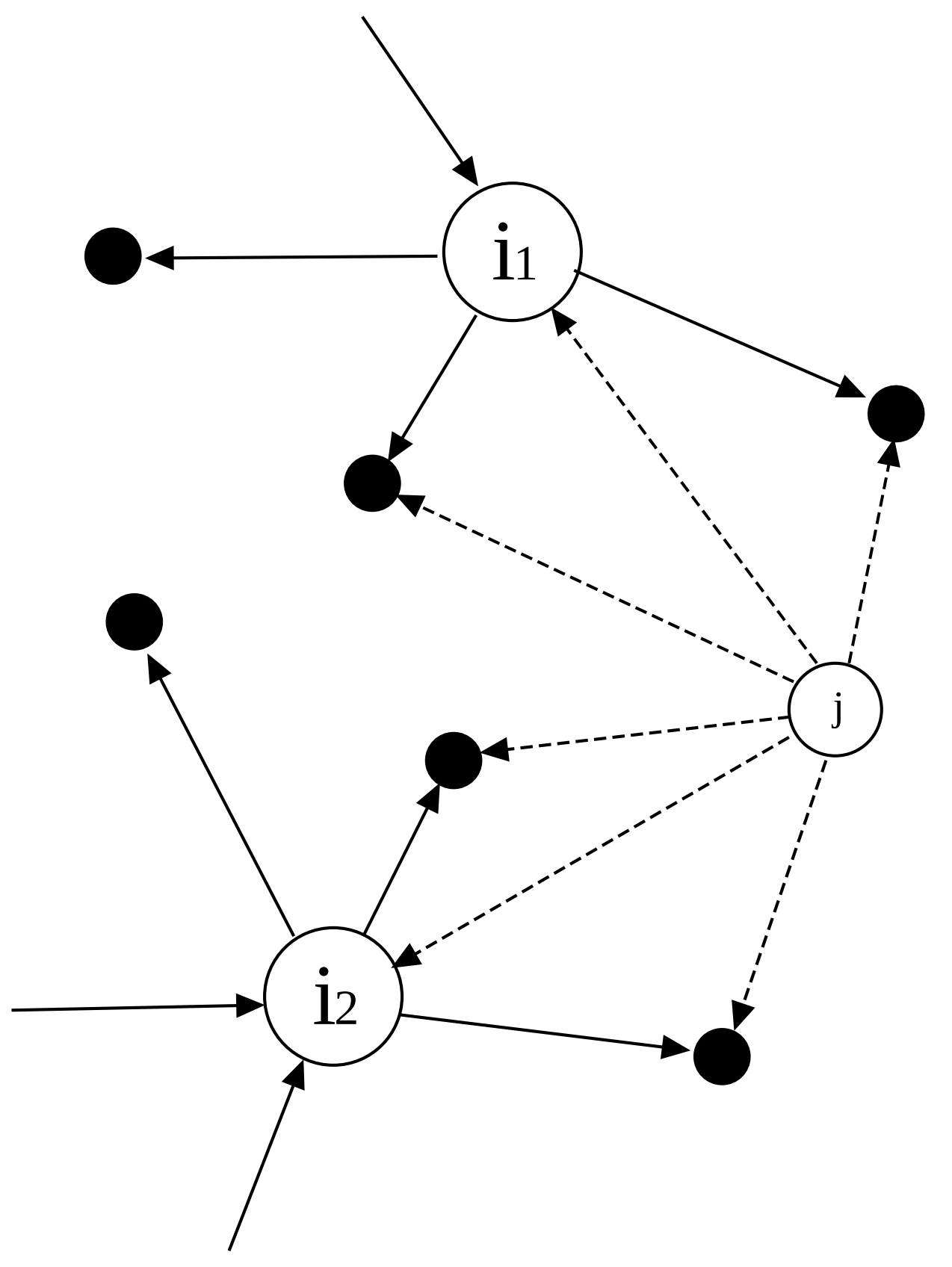
The archetypal example of a growing complex network process is the **Preferential Attachment** mechanism. Here nodes are introduced one by one and linked to old nodes in the network, randomly selected but with probabilities proportional to their degrees. The result is a network that exhibits a power-law degree distribution (called a scale-free network) which looks remarkably similar to many networked systems observed empirically, most famous perhaps is the network of scientific citations, where nodes represent articles and links represent the references between them.

Subsequent research suggests that subtly different mechanisms could be causing this behaviour: when a network grows in such a way that triads are created the result is a network which is not only scale-free, but also contains the levels of clustering expected in the systems they model. In this project we find our own approach to creating scale free clustered networks. The advantage of our method is that through careful selection of the parameters, the degree distributions of both the incoming and outgoing links and also the average clustering coefficient can be tuned to reach a wide range of values.

A toy model for clustered tunable networks

In each iteration,

- m ambassador nodes are randomly selected
- l descendants of each ambassador are also selected
- A new node is attached by a directed link to each of the selected nodes.



Example: The new node j attaches to $m = 2$ ambassador nodes found randomly in the network (i_1 and i_2) and $l = 2$ descendants of each one (the dark nodes). The out-degree of every node is $m(l+1)$.

Let N be the total number of nodes and N_k the number that have in-degree k. We find that as the network grows large, the probability that in one iteration a node i with degree k will be selected is

$$P(i;k) = \frac{m}{N}\left(1 + \frac{lk}{m(l+1))}\right).$$

The first term on the right hand side represents the probability of linking to i as an ambassador, the second is the probability of linking to i as a descendant of some other node. Using this we are able to write down a rate equation in the mean field

$$\frac{\partial N_k}{\partial N} = \frac{m}{N}\left[\left(1 + \frac{l(k-1)}{m(l+1)}\right)N_{k-1} - \left(1 + \frac{lk}{m(l+1)}\right)N_k\right] \tag{1}$$

and

$$\frac{\partial N_0}{\partial N} = 1 - \frac{m}{N}N_0. \tag{2}$$

We find $P_{\text{in}}(k)(= N_k/N)$ the proportion of nodes that have in-degree k for large N. By solving Eq.(1) and Eq.(2) we get

$$\left(\frac{l+1+m(l+1)}{l} + k\right)P_{\text{in}}(k) = \left(\frac{m(l+1)-l}{l} + k\right)P_{\text{in}}(k-1)$$

which has the exact mean field solution

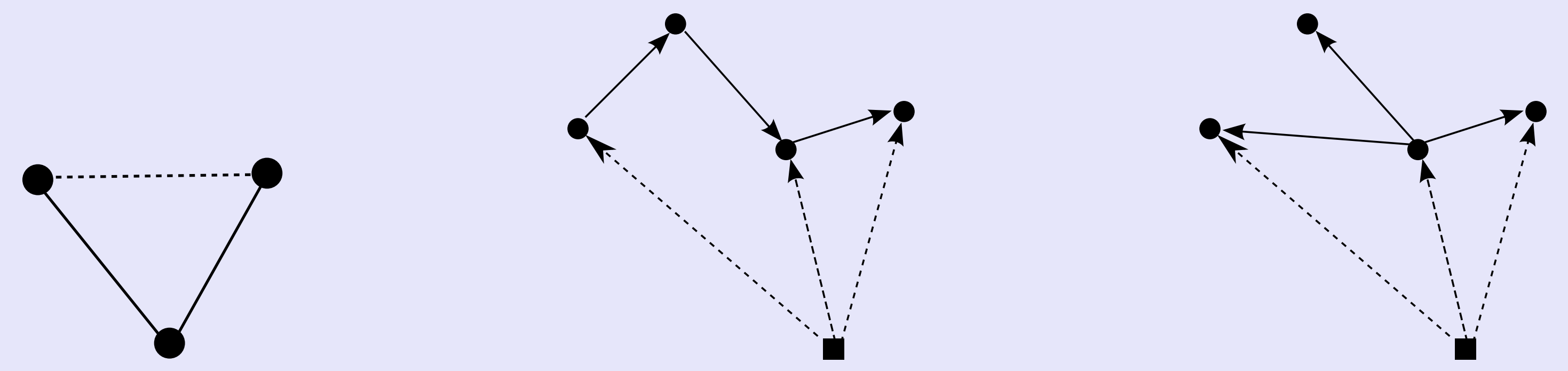
$$P_{\text{in}}(k) = \frac{1}{m+1}\frac{\Gamma[(2l+1+m(l+1))/l]\Gamma[k+m(l+1)/l]}{\Gamma[m(l+1)/l]\Gamma[k+(2l+1+m(l+1))/l]}.$$

The tail of the distribution follows a tunable power law

$$P_{\text{in}}(k) \sim k^{-\gamma} \text{ where } \gamma = \frac{2l+1}{l}.$$

Background: Network growth by triad formation

Growth mechanisms that produce highly clustered networks are well studied. They typically involve the creation of triads, i.e. connected triples of nodes in the network.



The most simple mechanism to achieve this is **Triadic Closure**, the process of adding links between nodes that share a common neighbour (left). This idea has interested sociologists since the early 20th century and has significance in *link prediction*, when combined with other growth mechanisms it forms a model for social networks [4]. Alternatively, triads can be formed when new nodes are introduced to the network. For example, by linking the new node to old ones selected via a **random walk** through the network (middle), or by a process of **link copying**; linking a new node to a random old node and to some of its neighbours (right). The link copying mechanism, first studied by Kumar et al, is often suggested as a suitable model for citation networks since authors of scientific papers tend to "copy" citations from the bibliographies of the other papers they cite [2, 3, 1]. A remarkable quality of these processes is that each one produces scale-free networks without any predetermined preferential linking, suggesting that creating clusters is, in at least some cases, the cause of scale-free topology seen in many interactive complex systems.

When m and l are random variables

In a *general model* we allow the values m and l to take a range of values. In each iteration: with probability q_m we select m ambassador nodes. Then for each of these, with probability p_l, we select l descendants (m and l are integers). The out degree can be written

$$P_{\text{out}}(s) = \sum_{n=1}^{\infty} q_n P\left(\sum_{i=1}^{n} x_i = s\right)$$

where the x_i are integer random variables that equal $l+1$ with probability p_l. To calculate the in-degree distribution we construct a rate equation from $P(i;k)$, the probability that the degree of a node i will increase by 1 during any iteration. If we assume the network is very large and ignore the effect of correlations we find

$$P(i;k) = \frac{\langle m\rangle}{N}(1+k\Phi) \text{ where } \Phi(p,q) = \sum_{l=1}^{\infty}\sum_{s=l}^{\infty}\frac{lp_l P_{\text{out}}(s)}{s}.$$

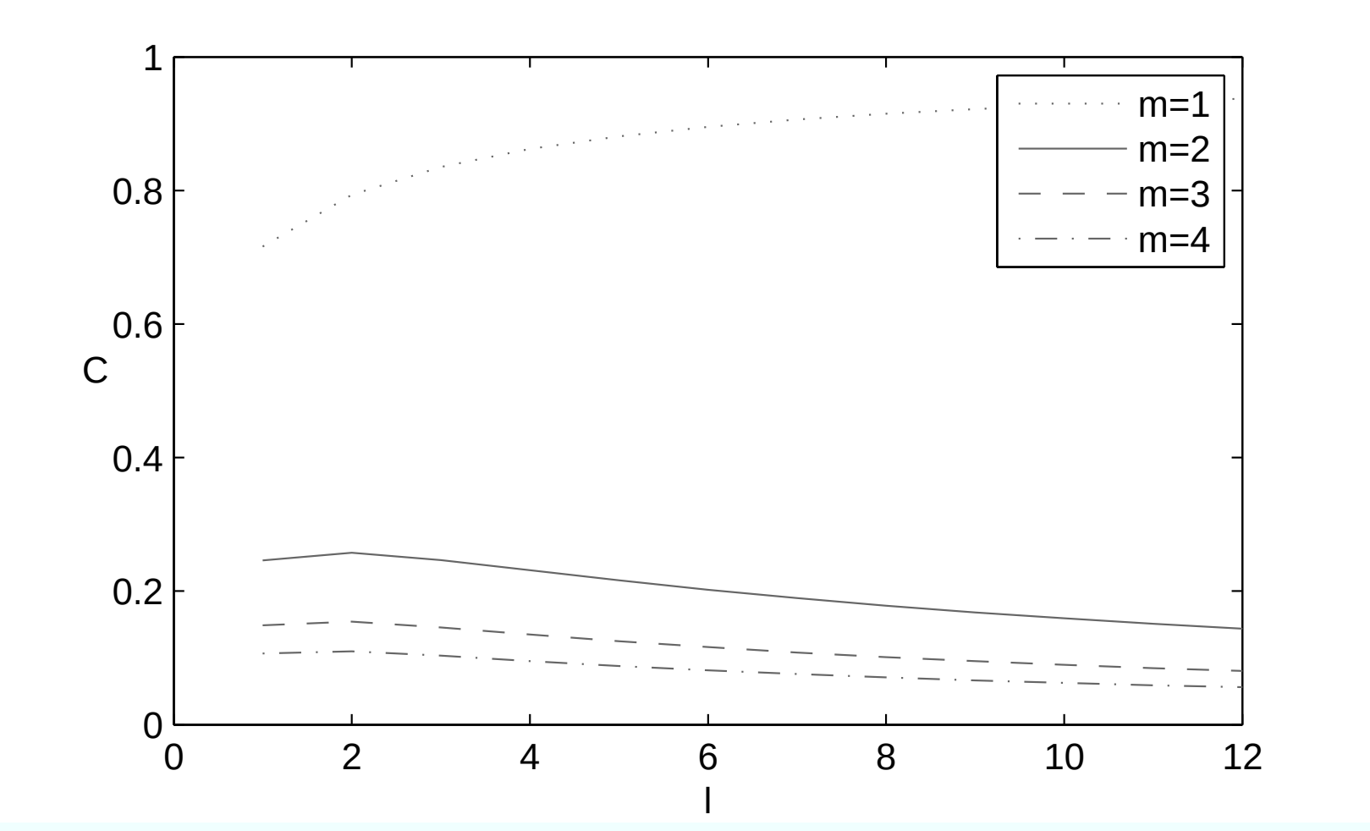
For large values of k, $P_{\text{in}}(k)$ has a power-law form

$$P_{\text{in}}(k) \sim k^{-\gamma} \text{ where } \gamma = 1 + \frac{1}{\langle m\rangle\Phi}.$$

Note that the power-law distribution is found for almost all choices of p_l and q_m.

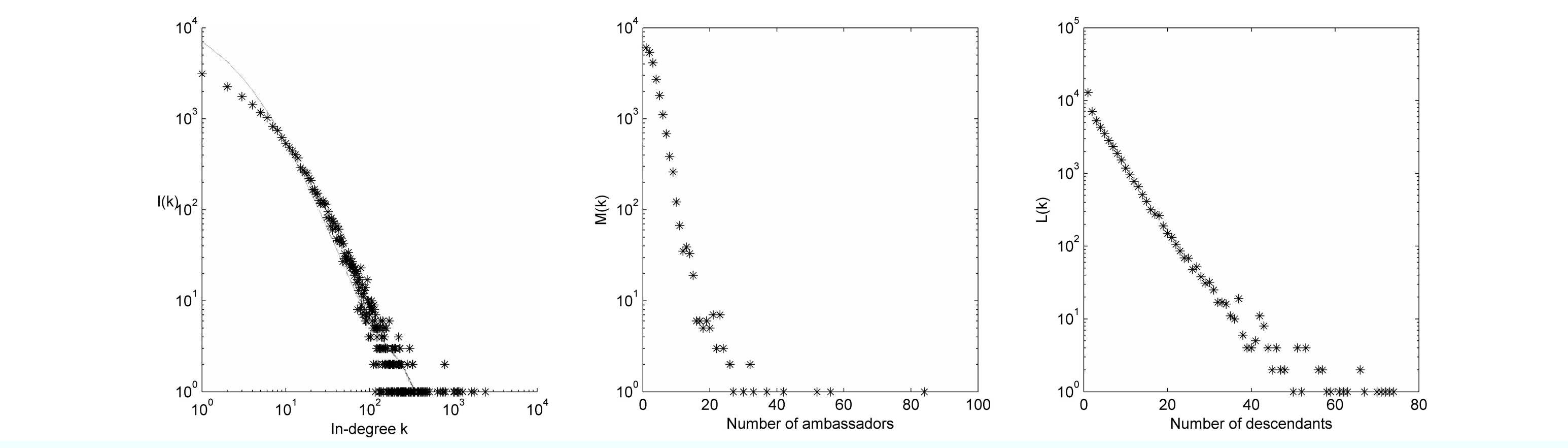
Tunable clustering

The clustering coefficient of a node i is defined as the proportion of pairs of neighbours of i that share a link. On the vertical axis we plot the average clustering coefficient over all nodes for the first 4 values of m in the *toy model*, against l on the horizontal axis. The clustering tends to zero as l grows large for all values of m with the exception of $m = 1$. In the *general model* it is easy to find parameter values to produce any clustering coefficient between 0 and 1.



The model vs. the data

Link copying mechanisms have been proposed as a model for citation networks. Our parametrization of this process allows the link creation in the model to behave similarly to the citing behaviour of authors.



The panel on the left shows the citation distribution of the **high energy physics** dataset. The dashed line shows the results of the model when the we use the distributions shown in the other two panels as parameters. These distributions are inferred from the data, $M(k)$ is equivalent to q_m and is the number of papers that link to k ambassador nodes. $L(k)$ is equivalent to p_l and is the number of occurrences of k papers being linked to from an ambassador.

Beyond citation networks

There are numerous potential applications for clustered network models beyond Scientometrics, similar analysis may be applicable in the following cases...

Twitter is a directed network of users "following" other users. A user may be directed to another user because they were "re-tweeted" by someone they already follow, an example of link copying.

The network of hyperlinks on Wikipedia respresents a semantic web of ideas spanning all human knowledge. Finding ways to measure and understand the cluster topology of this network is therefore an inportant task if we are to find efficient ways to navigate this web.

In recommender systems products are linked to each other if they have been bought by the same customer. Triads are formed and clustering increases when a customer buys a recommended product instead of a random one, this coould lead to an unfair power-law distribution of product sales.

With new technologies emerging that depend heavily on self organising networks, the need to gauge the intrinsic attractiveness of nodes is becoming ever more important. This research is at an early stage and we hope the work presented will provide a useful analytical foundation for such quantitative tools to be built on.

References

[1] E. Colman and G. Rodgers. Complex scale-free networks with tunable power-law exponent and clustering. *Physica A: Statistical Mechanics and its Applications*, 2013.

[2] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The web as a graph: Measurements, models, and methods. In *Computing and combinatorics*, pages 1–17. Springer, 1999.

[3] M. V. Simkin and V. P. Roychowdhury. Copied citations create renowned papers? *arXiv preprint cond-mat/0305150*, 2003.

[4] A. Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, 67(5):056104, 2003.

Acknowledgements

We are grateful to the EPSRC for their support.