Chapter 4

Chemical model diagnostics using graph theory and metrics.

"The complexities of cause and effect defy analysis."

- Douglas Adams, *Dirk Gently's Holistic Detective Agency*

4.1 Introduction

The node-link style structure has long been used to represent real-world relationships between items. Such a structure is complementary to our cognitive disposition towards pattern recognition [CITE]. It is for this reason that the node-link visualisation format has been used for anything ranging from transportation maps [CITE BECK] to the differentiation of ancestorial lineages of the human race, Figure 4.1. However, the abundance and complexity of real-world data often presents us with difficulties in manually representing it. One method to overcome this is through the use of computers to either analyse the data or make use of automated visualisation and graphing tools (i.e. Data-Visualisation) - both of which require a machine parseable representation of the data.

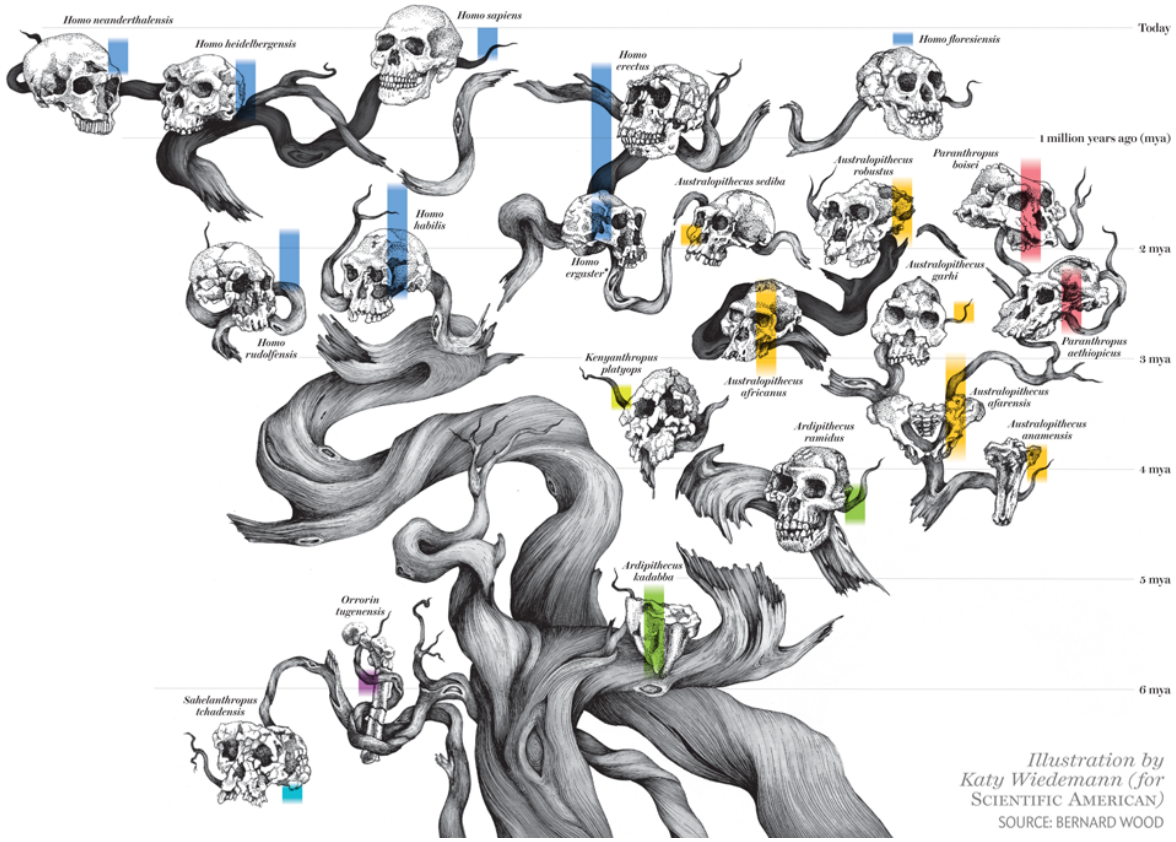


Figure 4.1: The human family tree showing the ancestral lineage to produce the modern-day human. Trees are a specific case example of the use of graph structure within visualisation. Source: [Wood, 2014]

In the field of mathematics a graph, $G(\nu, \epsilon, \omega)$, is defined as a function of items (vertices[1]), ν which are connected through a series of links (or edges[1]) representing any relationships between them, ϵ. Since relationships in the real world are rarely equivalent, we then encode the importance of each link in the form of an edge weight, or strength - ω. Such formats allow both numerical and computational

[1]The term node, item or vertex shall be used interchangeably for the remainder of this chapter. This also applies to links/relationships/edges and edge-weight/strength

algorithms to understand and interperate the graph structure, providing us with information about the data or make use of automated layout programs for visualisation (Chapter 2).

4.2 Graph Metrics

As our ability to gather more meaningful data increases, so do our datasets. This results with large, multivariate networks of inexplicable complexity. In such cases, our ability to draw out meaningful conclusions based on visualisation alone become highly inhibited. The solution once again falls to the field of mathematics, or more specifically Graph Theory through the use of a set of numerical algorithms or centrality metrics.

Centrality metrics are designed in such a way that they reveal important characteristics of the network and its underlying structure. They have been implemented for a range of contexts... the Russian spy network, roman etc (SEE BOOK ON DESK for examples), they reveal trends of ...

The mathematical nature of these allow us to alleviate some of the difficulties of large/complex graphs and may be used to complement or replace visualisations with regards to data mining. Their ability to identify important nodes, as well as their role within the network, makes them particularly useful for application on graph databases (neo4j [ref]). Here we may have billions of nodes representing anything from commodities to user names in trade, video streaming, finance or social media [referferfereferef].

4.2.1 Understanding Centrality - The Citation Graph

To demonstrate the use of such metrics we take the citation network created through scraping the results of google scholar [CITE] for all papers related to the Master Chemical Mechanism (MCM) [**?**].

The Master Chemical Mechanism

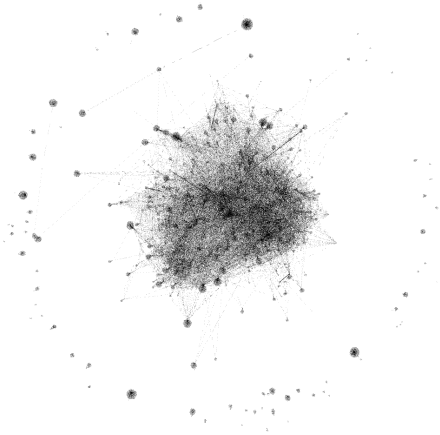
The master chemical mechanism is a near explicit representation of our foremost understanding of gas-phase tropospheric chemistry. The mechanism describes the oxidation of long-chain hydrocarbons and aromatics and the respective rates at which these occur. It has become an .... mechanism used to simulate or benchmark models and results in the field of atmospheric chemistry. Information on the chemistry, and how this can be used with regards to the following algorithms are presented in section 4.4

4.2.1.1 Collecting the data

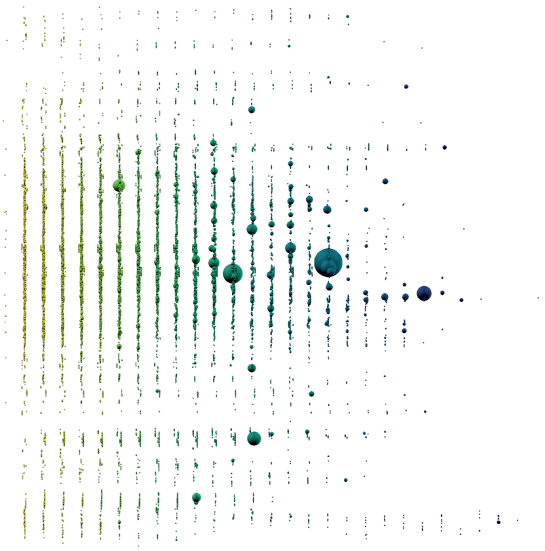
To obtain data regarding papers on the Master Chemical Mechanism, we begin by querying only articles containing the set of words: { *"Master", "Chemical", "Mechanism"* and *"MCM"* }. We then select the first 100 pages, each with 10 articles within them and collect their names. Following this, we repeat the process by looking at the top 1000 citations for each page. This results in a network of 15744 papers and 30178 citations[2]. Data collection was done using a tweaked version of *etudier* [Edsu and Ellis,].

We then created a three-dimensional visualisation tool using WebGL and THREE.js (REF). This allowed for the quick and efficient viewing, querying and interacting of the data. The additional dimension also helped identify the temporal change in the network by mapping node height relative to the year of publication.

[2]Note: this had the potential of returning up to 1000,000 nodes



(a) A 2D force directed representation of the network using the gephi software [Bastian et al., 2009]



(b) 3D orthographic camera view across the time axis - This shows papers with respect to publication date (yellow = recent.) Figure 4.2a

(c) 3D perspective camera view of the graph in Figure 4.2a

Figure 4.2: Still captures of 2D and 3D visualisations of the dataset. Node size corresponds to the number of citations, and colour (and z-axis) corresponds to the publication year for each paper.

4.2.1.2 Filtering the data

In the method used to web scrape data there exist several features which needed to be corrected. Firstly there exist papers within the dataset older than 1996 (when the first version of the MCM was created), Figure 4.2b. This is because papers, including the first MCM related papers, are likely to reference historic research and findings. Since we aim to explore the effect of introducing the master chemical mechanism, these are removed.

Next belt consisting of small satellite clusters of poorly connected clusters are apparent in Figure 4.2c. These are often papers which are often not within the field of atmospheric chemistry (thus the poor connectivity between main network body) but have cited mechanism related research, **??**. These will then also have a localised halo of papers citing them, due to the web scraping method used.

Finally, author names and some extended titles may be truncated with the use of ellipses. This is due to the web scraping script extracting these directly from the Google scholar page, and not the original articles themselves. It is worth noting that the results in this section are not explicit, but rather a demonstration of graph theory on a real-world dataset.

Fabrication of Bioinspired Actuated Nanostructures with Arbitrary Geometry and Stiffness	[Pokroy et al., 2009]
Temporal controls on dissolved organic matter and lignin biogeochemistry in a pristine tropical river	[Spencer et al., 2010]
Neuroproteomics in Neurotrauma	[Ottens et al., 2006]
Fast start-up of a pilot-scale deammonification sequencing batch reactor from an activated sludge inoculum	[Jeanningros et al., 2010]
Red blood cell oxidative stress impairs oxygen delivery and induces red blood cell aging	[Mohanty et al., 2014]
Life cycle assessment of Italian high quality milk production.	[Fantin et al., 2012]

Table 4.1: Example topics presented in some isolated satellite clusters.

Co-citation

The problem of removing clusters of adjunct research fields may be solved through the use of a co-citation graph. Such methods have been used to show forward propagating trends of papers in the same field, compared to the backwards propagation of citation graphs. Cocitation has been heavily used in the comparison of academic ... most notably in natures 150-year edition [CITE], where...

CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature Chaomei Chen

Using the methods above we reduce the citation graph to 451 papers and 2758 edges, and create an undirected co-citation graph of the same number of nodes, but 5402 edges.

4.2.1.3 Co-author

Finally, it is also useful to see what authors commonly publish together[3].

Using such a method we can build a co-author network which may draw attention to research areas/groups who perform work related to the MCM - for instance **??** hilights three main workgroup locations present in papers relating to the MCM. Additionaly as this is the simplest of the three networks above, it will also be used to visually show the properties of each centrality metric.

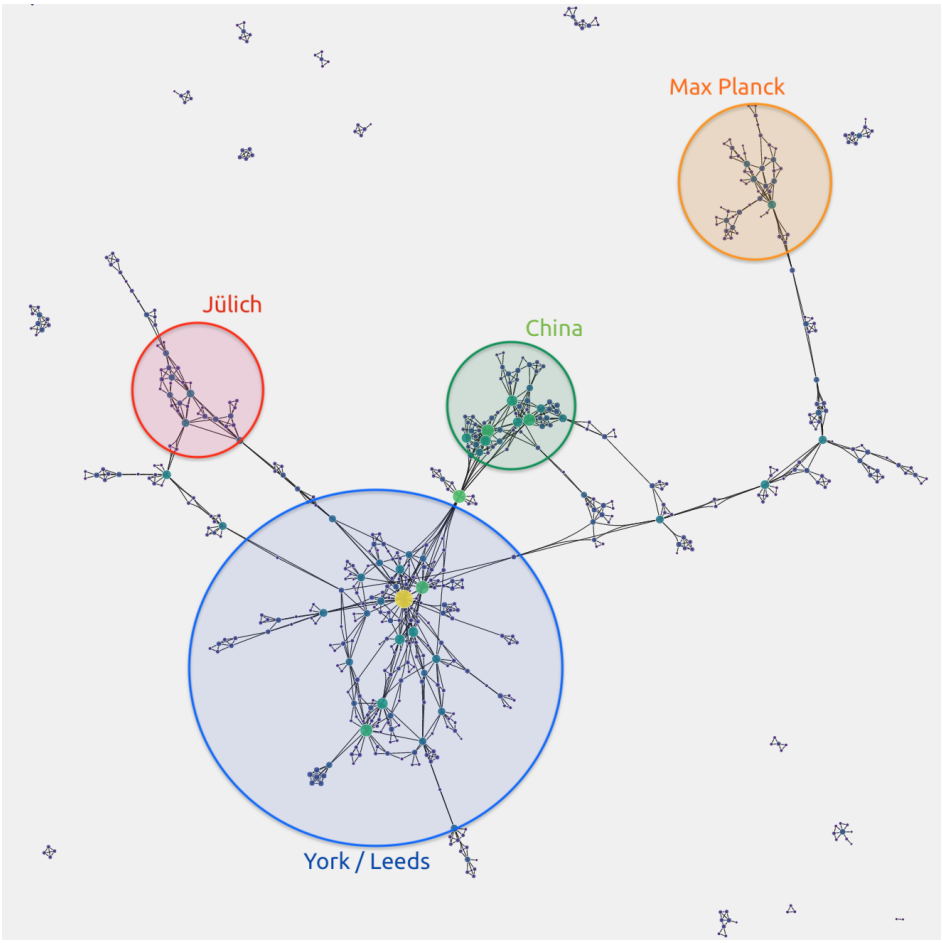


Figure 4.3: Workgroups within the co-author network.

[3]Disclaimer: as mentioned earlier, not all authors for every paper were recorded by the web scraping algorithm

4.2.2 Network analysis

This section aims to access the efficiency of graph centrality metrics in identifying important nodes within a network. To do this three inherently simple networks derived from the citing of papers related to the Master Chemical Model will be used. The first is the use of directed citation counts, the second using a co-citation network to find related topics, and the last a co-author map of the authors involved for each paper. The reasoning behind this is that if successful, the metrics should uncover groups of authors who work on similar topics, as well as papers which are important with regards to the MCM and the atmospheric community. Such types of analysis have been performed before as a means of predicting future citations and core papers within a field.

4.2.2.1 Degree

The simplest, and most intuitive, metric is degree centrality [Freeman, 1978]. This can be described as the sum of all links incident on a node - simply put, we count the number of citations going into and out of a paper. Such analysis has been used to determine influence in social media or in calculating the probability of a profile committing online auction fraud [Gemma, , Freeman, 1978].

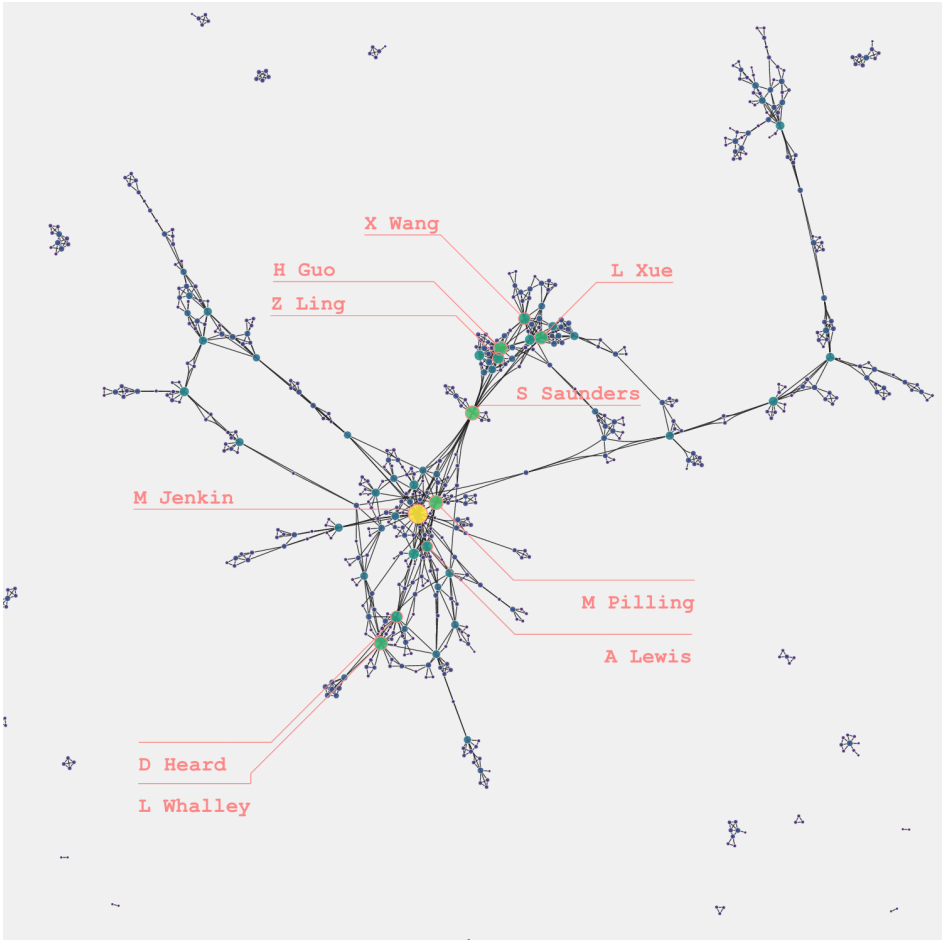


Figure 4.4: Degree centrality within the co-Author network

M Jenkin	39
S Saunders	25
M Pilling	25
H Guo	24
L Whalley	23
L Xue	22
D Heard	19
X Wang	19
Z Ling	18
A Lewis	17

Table 4.2: **Author network**: Top 10 ranked items using degree centrality

Protocol for the development of the Master Chemical Mechanism, MCM v3 Part A tropospheric degradation of nonaromatic volatile organic compounds	[?]	291
Protocol for the development of the Master Chemical Mechanism, MCM v3 Part B tropospheric degradation of aromatic volatile organic compounds	[?]	174
Development of a detailed chemical mechanism MCMv3. 1 for the atmospheric oxidation of aromatic hydrocarbons	[?]	103
Atmospheric oxidation capacity sustained by a tropical forest	[?]	60
Photochemical ozone creation potentials for organic compounds in northwest Europe calculated with a master chemical mechanism	[?]	54

Table 4.3: **Citation network**: Top 5 ranked items using degree centrality

Protocol for the development of the Master Chemical Mechanism, MCM v3 Part A tropospheric degradation of nonaromatic volatile organic compounds	[?]	340
Protocol for the development of the Master Chemical Mechanism, MCM v3 Part B tropospheric degradation of aromatic volatile organic compounds	[?]	250
Atmospheric oxidation capacity sustained by a tropical forest	[?]	187
Development of a detailed chemical mechanism MCMv3. 1 for the atmospheric oxidation of aromatic hydrocarbons	[?]	184
HO x radical regeneration in the oxidation of isoprene	[?]	176

Table 4.4: **Co-Citation network**: Top 5 ranked items using degree centrality

4.2.2.2 Directed Degree

In directed graphs, the direction of edges becomes important. To establish which nodes are important concerning the service provided by their edges, we may split their degree into inwards and outwards going links. This allows us to differentiate papers which cite many others, from those which are cited by others.

Since the only directed network within our set is the citation network, we shall use this for comparison with the normal degree.

In-Degree

Protocol for the development of the Master Chemical Mechanism, MCM v3 Part A tropospheric degradation of nonaromatic volatile organic compounds	[?]	0.63
Protocol for the development of the Master Chemical Mechanism, MCM v3 Part B tropospheric degradation of aromatic volatile organic compounds	[?]	0.38
Development of a detailed chemical mechanism MCMv3. 1 for the atmospheric oxidation of aromatic hydrocarbons	[?]	0.22
Photochemical ozone creation potentials for organic compounds in northwest Europe calculated with a master chemical mechanism	[?]	0.12
Atmospheric oxidation capacity sustained by a tropical forest	[?]	0.12

Table 4.5: **Citation network**: Top 5 ranked items using In-Degree centrality

Out-Degree

The MCM v3. 3.1 degradation scheme for isoprene	[?]	0.06
Atmospheric photochemical reactivity and ozone production at two sites in Hong Kong Application of a master chemical mechanismphotochemical box model	[?]	0.05
HOx budgets during HOxComp A case study of HOx chemistry under NOxlimited conditions	[?]	0.05
Modelmeasurement comparison of functional group abundance in pinene and 1,3,5trimethylbenzene secondary organic aerosol formation	[?]	0.05
Evaluation of and pinene degradation in the detailed tropospheric chemistry mechanism, MCM v3. 1, using environmental chamber data	[?]	0.05

Table 4.6: **Citation network**: Top 5 ranked items using Out-Degree centrality

4.2.2.3 Closness Centrality

Closeness centrality ranks all the nodes within a network on their ability to convey information to other nodes. We calculate the reciprocal of the sum of Dijkstra paths[4] to every other species [clo, , Sabidussi, 1966]. This creates a metric capable of representing the average distance of a node to every other within the network. Such a metric has been found useful in assessing the usefulness of intelligence gathering in terrorist networks, packet arrival in telecommunications and the calculation of word importance in key-phrase extraction [Krebs, 2002, Borgatti, 2005, Boudin,].

> ***Example analogy****If we take the UK rail network as an example, York station will have a high closeness value as it is well connected and central in location. This means it is easy to reach every other location when compared to other stations.*

[4]the shortest available path

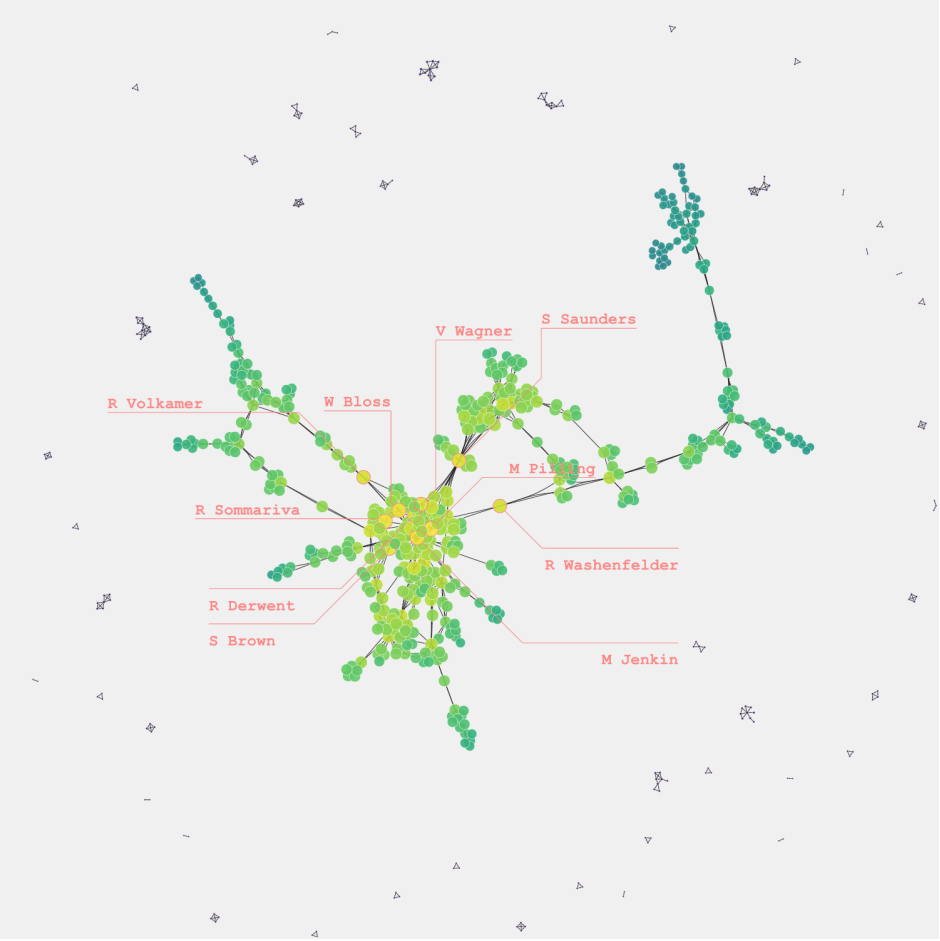


Figure 4.5: Closeness centrality within the co-Author network

M Pilling	0.149995
M Jenkin	0.146532
R Sommariva	0.145251
W Bloss	0.144052
S Brown	0.142059
S Saunders	0.140176
V Wagner	0.139281
R Derwent	0.136450
R Volkamer	0.136184
R Washenfelder	0.135918

Table 4.7: **Author network**: Top 10 ranked items using closeness centrality

Protocol for the development of the Master Chemical Mechanism, MCM v3 Part A tropospheric degradation of nonaromatic volatile organic compounds	[?]	0.67
Protocol for the development of the Master Chemical Mechanism, MCM v3 Part B tropospheric degradation of aromatic volatile organic compounds	[?]	0.53
Photochemical ozone creation potentials for organic compounds in northwest Europe calculated with a master chemical mechanism	[?]	0.45
World Wide Web site of a Master Chemical Mechanism MCM for use in tropospheric chemistry models	[?]	0.43
Photochemical ozone creation potentials for oxygenated volatile organic compounds sensitivity...	[?]	0.40

Table 4.8: **Citation network**: Top 5 ranked items using closeness centrality

Protocol for the development of the Master Chemical Mechanism, MCM v3 Part A tropospheric degradation of nonaromatic volatile organic compounds	[?]	0.80
Protocol for the development of the Master Chemical Mechanism, MCM v3 Part B tropospheric degradation of aromatic volatile organic compounds	[?]	0.68
Atmospheric oxidation capacity sustained by a tropical forest	[?]	0.61
Development of a detailed chemical mechanism MCMv3. 1 for the atmospheric oxidation of aromatic hydrocarbons	[?]	0.61
HO x radical regeneration in the oxidation of isoprene	[?]	0.60

Table 4.9: **Co-Citation network**: Top 5 ranked items using closeness centrality

4.2.2.4 Betweenness

In social networks, it is often important not only to know who has the greatest reach (closeness centrality) but also where bottlenecks or 'broker' positions occur. The betweenness centrality is a count of the number of geodesics (shortest) paths that pass through a specific node in the network[5]. Nodes with a high betweenness since they control, or limit, the amount of information that can be transferred across the network. If a node lies on the shortest path between two other nodes, we may consider it a 'pivotal' node [Needham and Hodler, 2019]. should such a node then be removed, we incur either a deviation, whereupon a longer path is required, or two separate connected subgraphs [Freeman et al., 1991, Freeman, 1977, Brandes, 2001, Borgatti, 2005].

Example analogy *Expanding on the UK rail network analogy, Shrewsbury station serves the critical role of connecting many lines from England to Wales. In removing this station, routes from the Liverpool or Manchester to Cardiff will be greatly increased. Additionally, the Aberystwyth section of the line will then become isolated from the rest of the country.*

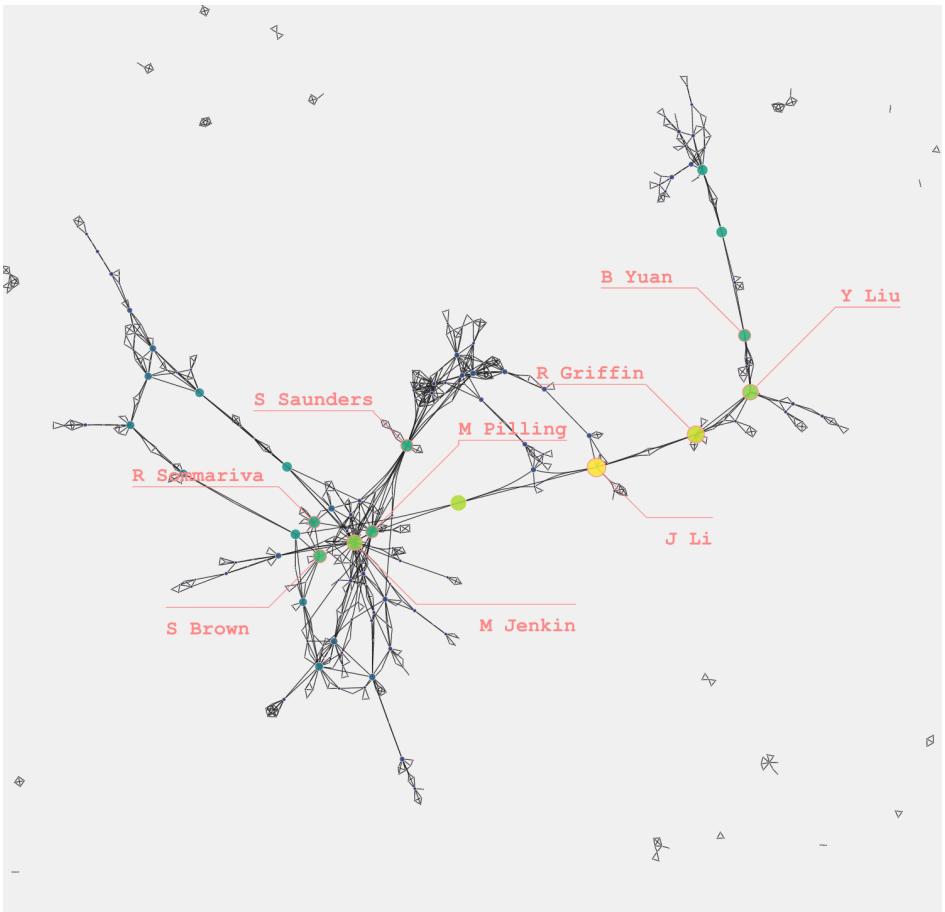


Figure 4.6: Betweenness centrality within the co-Author network

[5]In cases where there are multiple possible 'shortest' paths, we can account for this using the denominator

J Li	0.180998
R Griffin	0.162558
R Washenfelder	0.153024
Y Liu	0.142194
M Jenkin	0.139818
S Brown	0.110188
M Pilling	0.102816
B Yuan	0.099914
S Saunders	0.097255
R Sommariva	0.094757

Table 4.10: **Author network**: Top 10 ranked items using betweenness centrality

Impacts of mechanistic changes on HOx formation and recycling in the oxidation of isoprene	[?]	0.01
Protocol for the development of the Master Chemical Mechanism, MCM v3 Part A tropospheric degradation of nonaromatic volatile organic compounds	[?]	0.01
The regional atmospheric chemistry mechanism, version 2	[?]	0.01
Evaluation of and pinene degradation in the detailed tropospheric chemistry mechanism, MCM v3. 1, using environmental chamber data	[?]	0.01
A review of tropospheric atmospheric chemistry and gasphase chemical mechanisms for air quality modeling	[?]	0.01

Table 4.11: **Citation network**: Top 5 ranked items using betweenness centrality

Protocol for the development of the Master Chemical Mechanism, MCM v3 Part A tropospheric degradation of nonaromatic volatile organic compounds	[?]	0.12
Protocol for the development of the Master Chemical Mechanism, MCM v3 Part B tropospheric degradation of aromatic volatile organic compounds	[?]	0.10
HO x radical regeneration in the oxidation of isoprene	[?]	0.05
The MCM v3. 3.1 degradation scheme for isoprene	[?]	0.04
Development of a detailed chemical mechanism MCMv3. 1 for the atmospheric oxidation of aromatic hydrocarbons	[?]	0.04

Table 4.12: **Co-Citation network**: Top 5 ranked items using betweenness centrality

4.2.2.5 Spectral methods and matrix analysis

Graphs can often be represented in the form of relationship (adjacency) matrixes (ref Chapter 1). This allows us to apply the theory of linear maps, such as eigenvectors and values, to stochiometric data in matrix form. Such methods have been around since the 1950s, [R. Seeley, 1949], but mainly became popular with the release of Larry Page's page-rank algorithm [Page et al., 1999] - the algorithm that began google. These methods, in addition to the HITS algorithm subsubsection 4.2.2.6, make use of a graphs native matrix representation to calculate node importance. Spectral algorithms can be broken down into four categories [Vigna, 2016]:

	No Normalisation	Row Normalisation
No Damping	Eigenvector [Bonacich, 1987, Bonacich, 2007]	Markov Chain Steady State [R. Seeley, 1949]
Damping	Katz [Goh et al., 2001]	Total Effect Centrality PageRank [Page et al., 1999]

Here damping terms represent the probability of moving to new random starting position, allowing for the user to 'randomly select a new webpage' or leave an isolated cluster. Normalisation of the matrix does not affect the node ranking, but merely adjusts the numerical output of the algorithm. It is for this reason that its overall practicality may be debated [Vigna, 2016]. Since page rank is

the most common of these methods and allows for a tune-able degree of randomness within network propagation, this shall be discussed in more detail.

4.2.2.6 Hypertext Induced Topic Search (HITS)

One eigenvector based algorithm with use in classifying webpages into Hubs and Authorities is HITS [Kleinberg, 1999, Langville and Meyer, 2005, Kumar and Upfal,]. This is a more complex method of looking at the role of a node as an information provider (lots of outward links), i.e. an Authority or an information receiver - a Hub. Although there are many similarities to the directed degree centrality, this algorithm explores how information is propagated across the whole system, and can, therefore, give slightly different results.

A Common Representative Intermediates CRI mechanism for VOC degradation. Part 3 Development of a secondary organic aerosol module	[?]	0.01
HOx budgets during HOxComp A case study of HOx chemistry under NOxlimited conditions	[?]	0.01
Detailed chemical analysis of regionalscale air pollution in western Portugal using an adapted version of MCM v3. 1	[?]	0.01
Box model studies of the secondary organic aerosol formation under different HCNOx conditions using the subset of the Master Chemical Mechanism for pinene	[?]	0.01
Reporting the sensitivity of laserinduced fluorescence instruments used for HO2 detection to an interference from RO2 radicals and introducing a novel approach that	[?]	0.01

Table 4.13: **Citation network**: Top 5 ranked items using Hub centrality

Protocol for the development of the Master Chemical Mechanism, MCM v3 Part A tropospheric degradation of nonaromatic volatile organic compounds	[?]	0.11
Protocol for the development of the Master Chemical Mechanism, MCM v3 Part B tropospheric degradation of aromatic volatile organic compounds	[?]	0.07
Development of a detailed chemical mechanism MCMv3. 1 for the atmospheric oxidation of aromatic hydrocarbons	[?]	0.04
Atmospheric oxidation capacity sustained by a tropical forest	[?]	0.02
Photochemical ozone creation potentials for organic compounds in northwest Europe calculated with a master chemical mechanism	[?]	0.02

Table 4.14: **Citation network**: Top 5 ranked items using Authority centrality

Page Rank

Page rank is the best known of all centrality algorithms [Needham and Hodler, 2019]. As with all eigenvector methods, it measures the transitive influence in nodes, taking the effect of neighbours, and by proxy, their neighbours into account. In the context of web pages or citations, a link from a highly ranked, or credible, source holds more weight than one from someone less credible.

History of page rank, app web pages citation counts. biophysics, etc, etc,..

The Google Matrix

To solve for page rank, one must first construct the google matrix. Once this has been done, iterations of the power method can be applied until convergence is reached.

Building the google matrix begins by turning our graph adjacency matrix $A_{i,j}$ into a Markov matrix $M_{i.j}$. The simplest way is to take our dyadic link map **??**, and divide each column j by the sum of the total outgoing links of node j, Algorithm 1. Dangling nodes are species with no outgoing links. In chemical mechanisms, these are generally removed but could represent sinks within a system. In the case of dangling nodes, either a personalised list of values or a constant value, $1/n$, replaces the

zero columns[6]. This construction results in a normalised[7] matrix of Markov chains representing the fractional production for node j from all other nodes.

Algorithm 1 Adjacency to Markov matrix.

```
Obtain graph adjacency matrix, A_{i,j}.
repeat
    for each j ∈ columns do
        M(:, j) ← A(:, j)/Σ_{i=1,n} A(j, i)
    end for
until Σ_{i=1,n} M(i, j) = 1
```

The google matrix can now be defined using Equation 4.1. Cyclic reactions and nodes that only point towards each other within a group can 'trap' the user, increasing their ranks. A damping factor, β, can be used to reduce this through selecting a probability that a user follows an existing link, typically $\beta = 0.85$, and a probability that they randomly select another page[8], $(1-\beta)$. The value of β will vary with application - a study in the application to biological data found an optimum value of $\beta = 0.694$ using Bayesian analysis [Hobson et al., 2018], however in most cases the typical value will suffice.

$$G_{i,j} = \beta M + \frac{1-\beta}{n} \tag{4.1}$$

β - *Probability the user follows a link*

$(1-\beta)$ - *Probability the user does not follow a link (teleportation)*

n - *Number of items / species*

M - *Normalised markov matrix*

Solving the algebra

Once defined, the google matrix can be solved by propagating a ones vector, r of length n, where n is the number of species using Algorithm 2. This is repeated until a pre-defined tolerance, ϵ is reached. For best results, this can be set to the precision of the program.

For smaller systems, it is possible to use lapack [lap,], linear algebra solvers such as those used by numpy [Oliphant, 2006]. However, if a network is large, computing an $n \times n$ matrix may be very memory consuming. It is then possible to apply the methods as described above using a sparse matrix

[6] Where n is the number of nodes

[7] $\Sigma_{i=1,n} M(i,j) =$ unity

[8] Also known as teleportation.

Algorithm 2 Solving the google matrix linear algebra

1: Define value vectors $\bar{r}_t$ and $\bar{r}_{t+1}$:
2: $\bar{r}_t = [1_1, 1_2, ..., 1_n]$, $\bar{r}_{t+1} = [0_1, 0_2, ..., 0_n]$
3:
4: **while** $||\bar{r}_{t+1} - \bar{r}_t|| > \epsilon$ **do**
5: $\bar{r}_{t+1} \leftarrow M.\bar{r}_t$
6: $\bar{r}_t = \bar{r}_{t+1}$
7: **end while**

on a per-node bases [Jones et al., 01 , Hagberg et al., 2008]. A comparison of the page rank algorithm is seen in **??**.

Prediction

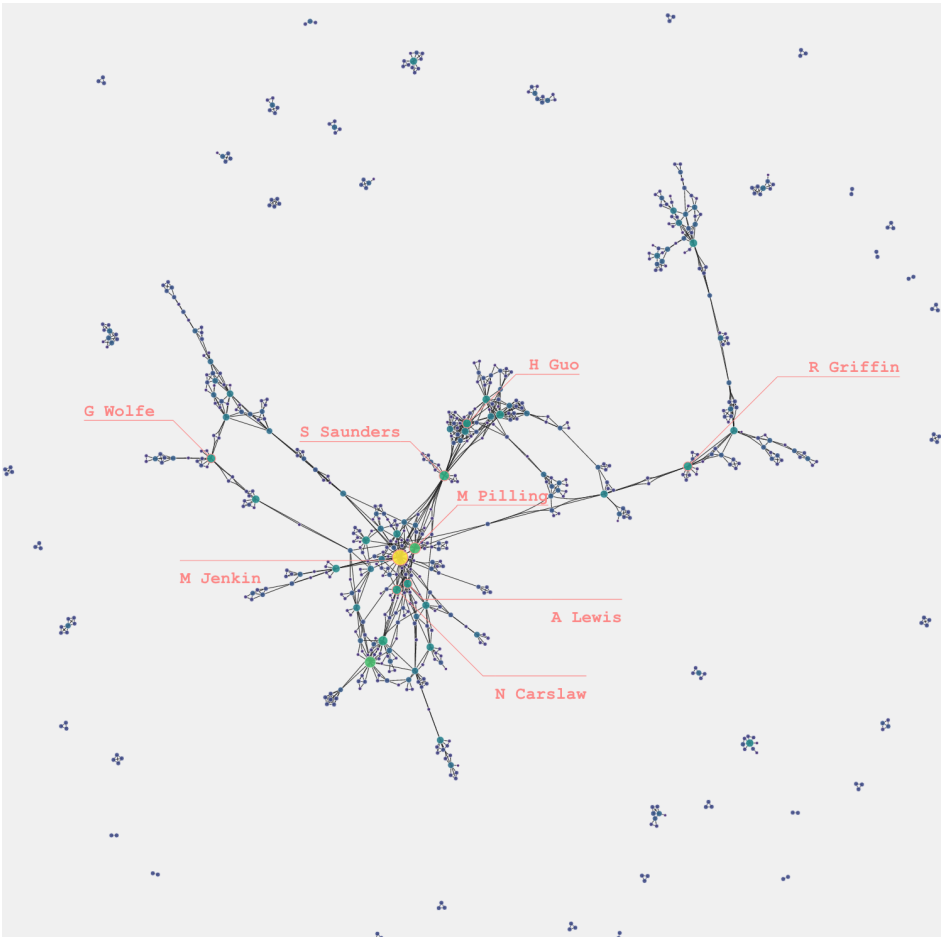


Figure 4.7: Page Rank centrality within the co-Author network

M Jenkin	0.010435
L Whalley	0.006589
M Pilling	0.006488
S Saunders	0.005591
D Heard	0.005192
N Carslaw	0.004833
H Guo	0.004594
G Wolfe	0.004523
A Lewis	0.004508
R Griffin	0.004500

Table 4.15: **Author network**: Top 10 ranked items using pagerank centrality

World Wide Web site of a Master Chemical Mechanism MCM for use in tropospheric chemistry models	[?]	0.09
Protocol for the development of the Master Chemical Mechanism, MCM v3 Part A tropospheric degradation of nonaromatic volatile organic compounds	[?]	0.07
Photochemical ozone creation potentials for organic compounds in northwest Europe calculated with a master chemical mechanism	[?]	0.07
Protocol for the development of the Master Chemical Mechanism, MCM v3 Part B tropospheric degradation of aromatic volatile organic compounds	[?]	0.05
Modeling OH, HO2, and RO2 radicals in the marine boundary layer 2. Mechanism reduction and...	[?]	0.03

Table 4.16: **Citation network**: Top 5 ranked items using pagerank centrality

Protocol for the development of the Master Chemical Mechanism, MCM v3 Part A tropospheric degradation of nonaromatic volatile organic compounds	[?]	0.06
Protocol for the development of the Master Chemical Mechanism, MCM v3 Part B tropospheric degradation of aromatic volatile organic compounds	[?]	0.03
Atmospheric oxidation capacity sustained by a tropical forest	[?]	0.03
Development of a detailed chemical mechanism MCMv3. 1 for the atmospheric oxidation of aromatic hydrocarbons	[?]	0.02
HO x radical regeneration in the oxidation of isoprene	[?]	0.02

Table 4.17: **Co-Citation network**: Top 5 ranked items using pagerank centrality

4.3 What kind of graph is the MCM?

Graph theory allows the mathematical evaluation and data mining of relational data. It can be used to extract and evaluate features existing in nodes, connections or entire graphs against each other. In this section, we shall apply graph theory to the entire MCM network, and determine any characteristics which define it. To do this we extract several hundred subsets and explore how the network scales.

4.3.1 Network density

The density of a network is defined as the ratio of edges against the total number of possible edges for a complete graph of the same size. This allows us to determine the sparsity of the network and determine if adding more chemistry saturates the graph. Figure 4.8 reveals that in the addition of more species (or nodes), the total number of edges (or reactions) decreases. This is because, in the addition of larger species, we introduce new branches of chemistry which need to be oxidised and broken up before they can react with many other species. Since these branches are somewhat isolated from the rest of the chemistry, they decrease the network density, even though their addition may increase the amount of chemistry that occurs within it.

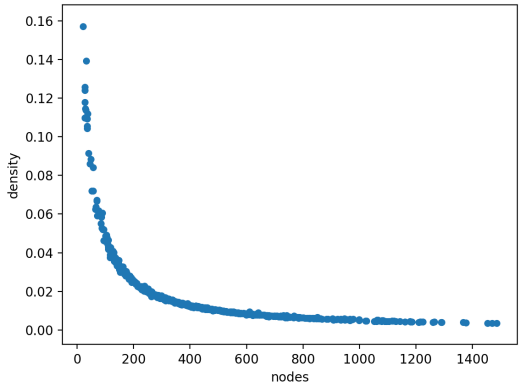


Figure 4.8: Exploring how the MCM graph density scales with number of species.

4.3.2 Small world Phenomena

The small world phenomena is highly common in biological or social networks, [Watts and Strogatz, 1998]. Such networks tend to have high local clustering (cliques) and a short path length between elements [Humphries and Gurney, 2008]. This makes it easy to reach all parts of a network with only a couple of hops or reactions, and is commonly known as the 'six degrees of separation'.

One method for establishing the small world-ness of a graph is through the calculation of the ω coefficient:

$$\omega = L_r/L - C/C_l \tag{4.2}$$

Where C and L are the average clustering coefficient and the shortest path length of the graph. The results are then compared using the average shortest path length, L_R, and clustering coefficient C_l for an equivalent random and lattice graph. This gives a result between [-1,1], with 0 exhibiting perfect small world-ness.

In [Jacob and Lapkin, 2018] it was found that the Reaxys chemical space exhibited some small world behaviours. Analysis for several MCM subsets, Figure 4.10, shows very small mechanisms (e.g. methane, ethane only) tend to have a lattice-like appearance, with larger more complex exhibiting small-world features, on the side of a random graph. This reflects the idea of tropospheric chemistry undergoing a series of localised reaction before oxidising sufficiently to connect to the rest of the network [REF]

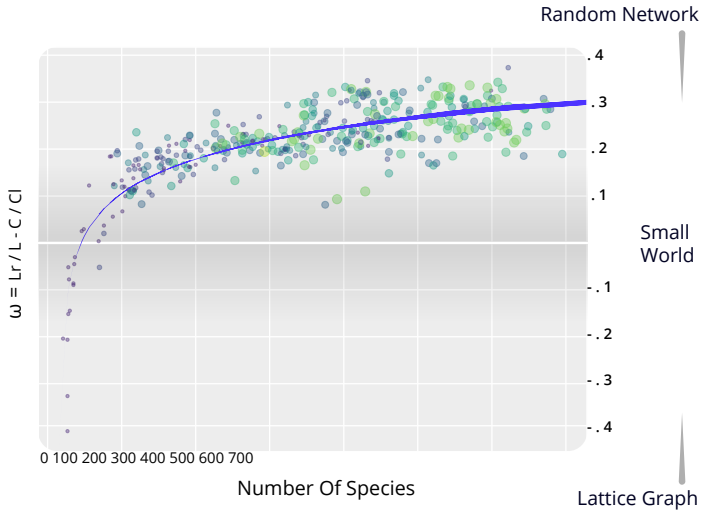


Figure 4.9: Small world-ness with a range of MCM subsets. Node size represents the number of reactions.

4.3.3 Power Law and Scale-free graphs

A scale-free graph is listed where the ratio between nodes and links follows a power-law distribution. This produces a hub and spoke architecture, and is commonly seen in graphs representing connections between websites on the internet.

(paper on how many power-law graphs are exponential or not power law.,)

Kolomogorov-Smirnov distance

The kolomogorov Smirnov distance is a method of comparing data to a distribution ...

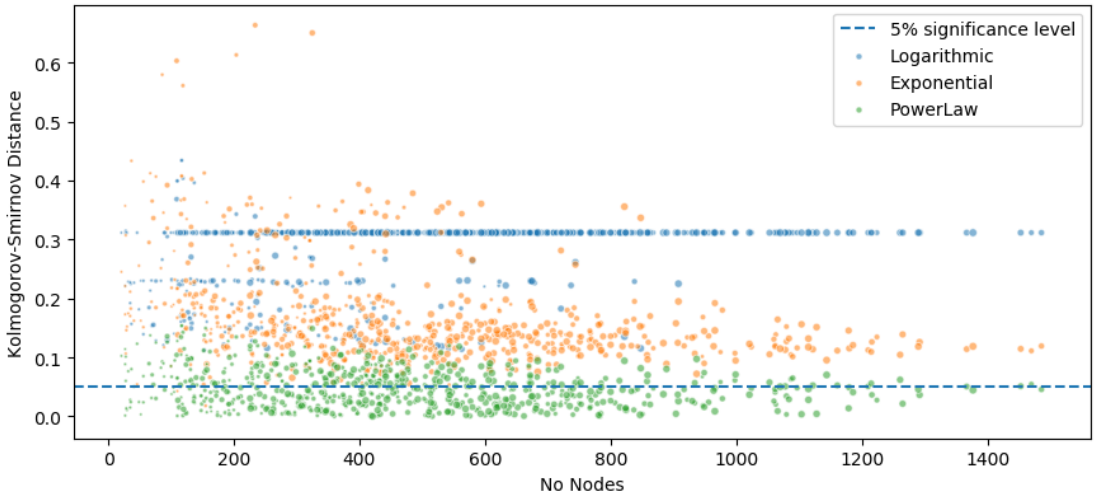


Figure 4.10: Comparing the MCM subsets against a power law and exponential distribution

4.4 A weighted Case study

So far we have only looked at the effect of graph metrics on the network structure. When we run a mechanism through a model, we not only have directed chemistry but the rates that concentration changes and reactions occur play a large part in how the chemistry of a system evolves. To diagnose the models and how the data within them changes, we have to add information on the strength of each relationship. To decide upon to the best way to do this we first view the existing methods of model analysis.

4.4.1 Existing methods of analysis

4.4.1.1 Concentration time series

The concentration of a species tells us information about its abundance within the atmosphere. As time evolves, species are oxidised, bonds are made/broken and the overall composition and ratios of concentrations of species change. In examining a species concentration we cannot only map its changes throughout time but also establish how changes in the initial conditions can affect the result of a simulation. We can then use this in predicting how changes in policy or management may influence the atmosphere. Additionally we can also determine influential factors and relationships, for example the inverse response between NO_2 and NO in Figure 4.11.

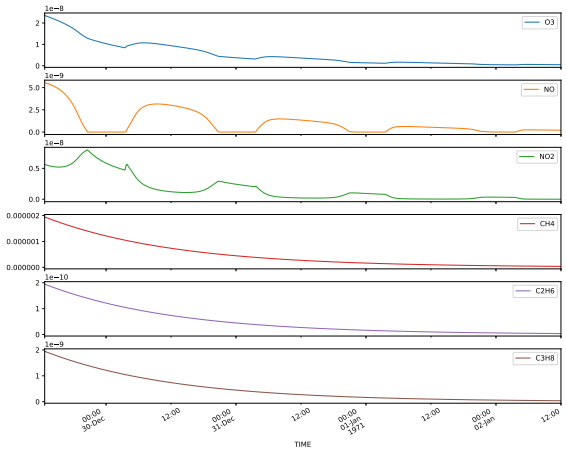


Figure 4.11: Concentration time series: The classic method for identifying changes in a model. This multi-plot shows the changes in concentration profiles for all initialised species following an initial spin-up to steady state.

4.4.1.2 Rate of Production and Loss

Analysing the concentration allows the comparison of a series of scenarios or runs change concerning their initial conditions and simulation length. Although these can tell us how, and how much, each species changes over time it does not rank or quantify the specific reactions to which this may be attributed. Rate of Production [9] Analysis (ROPA) provides a method for establishing the total contribution from each reaction by calculating the change of concentration (concerning time) for the produced species - the Flux.

$$r_1 = \eta A + \omega B \xrightarrow{\kappa_1} C \qquad \text{Reaction 1: ODE} \tag{4.3}$$

$$\frac{\delta C}{\delta t} = \Gamma_{r_1} = [A][B]\ \eta\omega \times \kappa_1 \qquad \text{Flux } (\Gamma) \tag{4.4}$$

Taking Figure 4.12 as an example, the ROPA plot allows us to rank the most influential reactions in the production of CH_3CO_3. In complex multi-body reactions isolating the exact cause of concentration change may prove difficult. Cyclic reactions, such as PAN $\longrightarrow$ CH_3CO_3 + NO_2 and CH_3CO_3 + NO_2 $\longrightarrow$ PAN, suggest a high importance to the production and loss of CH_3CO_3. However, with both reactions of similar magnitude and opposing directions the net effect is only marginal. To account for this we can calculate the individual contribution of one species on another using the Jacobian method.

[9] and loss

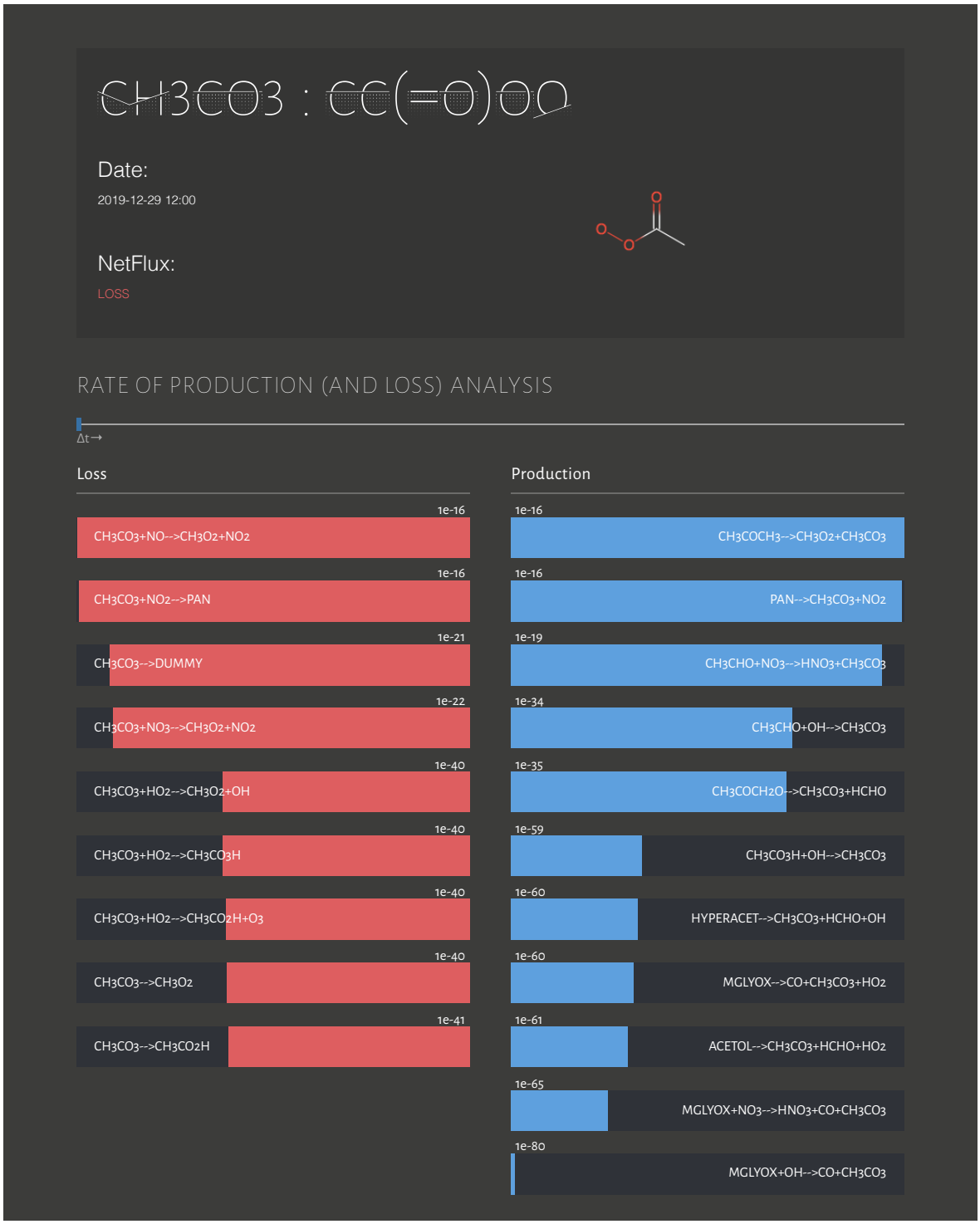


Figure 4.12: Rate of production and loss analysis plot for CH_3CO_3 exhibiting a net loss (daytime)

4.4.1.3 Analysing the Jacobian

The jacobian is a square matrix of partial derivatives. In taking the gradient of each reaction we can determine the individual (partial) contribution one species has on another. Since the jacobian is the rate of change of the flux, it may be described as the acceleration of our system, with the flux representing the velocity and the concentration - the displacement. This further bit of mathematics allows us to calculate how much an effect changing a species by 1% has on other species within the model at that timestep. It is for this reason that it forms the basis for many of the connectivity-based sensitivity and reduction calculations [REF connectivity method].

Calculating the Jacobian Matrix

To calculate the contribution of each species we start by differentiating the flux to produce a series of partial derivatives Equation 4.7 , Equation 4.6.

$$\frac{\partial}{\partial A} \cdot \frac{\partial C_{r_1}}{\partial t} = \eta\omega B\kappa_1 \qquad \Gamma \text{ influence from A} \tag{4.5}$$

$$\frac{\partial}{\partial B} \cdot \frac{\partial C_{r_1}}{\partial t} = \eta\omega A\kappa_1 \qquad \Gamma \text{ influence from B} \tag{4.6}$$

These may then be aggregated to produce a collection of partial derivatives from all reactions containing our two species A and B Equation 4.7.

$$\mathbf{J}_{A,B} = \frac{\partial f_A}{\partial B} = \frac{\partial}{\partial B} \cdot \left(\frac{\partial \Sigma_{r_1}}{\partial t} + \frac{\partial \Sigma_{r_2}}{\partial t} + \cdots + \frac{\partial \Sigma_{r_n}}{\partial t} \right) \tag{4.7}$$

With this methodology, it is then possible to construct a square matrix of all first-order partial derivatives (the Jacobian) for all species.

$$\mathbf{J}_{i,j} = \begin{bmatrix} \frac{\partial f_1}{\partial v_1} & \frac{\partial f_1}{\partial v_2} & \cdots & \frac{\partial f_1}{\partial v_n} \\ \frac{\partial f_2}{\partial v_1} & \frac{\partial f_2}{\partial v_2} & \cdots & \frac{\partial f_2}{\partial v_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial v_1} & \frac{\partial f_n}{\partial v_2} & \cdots & \frac{\partial f_n}{\partial v_n} \end{bmatrix}_{i,j=1}^{n,n}$$

4.4.2 Graph construction methodology for simulated data

4.4.2.1 Flux Based Graphs

For models which for which only the calculated rates are provided, it is possible to calculate the flux at each timestep using Equation 4.4. A multi-graph[10] joining each reactant to each product can then be constructed with each edge weighted by the amount provided by the flux calculation. For many cases, this will be sufficient and allow for the correct results.

4.4.2.2 Jacobian Graphs

Most atmospheric models combine the partial flux contributions from each reaction to a single relational matrix. This is known as the Jacobian matrix (discussed above). The reason for this is that it greatly reduces the number of explicit calculations required by each step by combining them into an $n \times n$ matrix, where n is the number of species. For models capable of returning this, or anyone with the required skills to extract this each timestep from the model runs, it can provide not only a simplified graph of the reactions but also a truer representation of the relationships between species than the Flux based method. For the remainder of the work an updated version of the Dynamically Simple Model of Atmospheric Chemical Complexity (DSMACC), [**?**] is used [**?**]. Here we edit the kinetic pre-processor output, [**?**] to release the values of the Jacobian Matrix.

A practical Example using the MCM

Taking a single equation, Equation 4.8, from the MCM we may calculate the jacobian relationships between species and convert them into a graph.

[10] A graph with multiple edges between nodes

$$C_2H_6 + OH \xrightarrow{\kappa_1} C_2H_5O_2 \tag{4.8}$$

For simplicity in this example, this will be the only equation for our mechanism. The resultant Flux Equation 4.9 and resultant Jacobian Equation 4.10 may be calculated.

$$\Gamma = [C_2H_6][OH]\kappa_1 \tag{4.9}$$

$$\mathbf{J}_{i,j} = \begin{bmatrix} \frac{\partial f_{[C_2H_6]}}{\partial t\ \partial[C_2H_6]} & \frac{\partial f_{[C_2H_6]}}{\partial[OH]} & \frac{\partial f_{[C_2H_6]}}{t\partial[C_2H_5O_2]} \\ \frac{\partial f_{[OH]}}{\partial t\ \partial[C_2H_6]} & \frac{\partial f_{[OH]}}{\partial t\ \partial[OH]} & \frac{\partial f_{[OH]}}{\partial t\ \partial[C_2H_5O_2]} \\ \frac{\partial f_{[C_2H_5O_2]}}{\partial t\ \partial[C_2H_6]} & \frac{\partial f_{[C_2H_5O_2]}}{\partial t\ \partial[OH]} & \frac{\partial f_{[C_2H_5O_2]}}{\partial t\ \partial[C_2H_5O_2]} \end{bmatrix}_{i,j=1}^{3,3} \tag{4.10}$$

Since not all species react with all other species, we can remove reactions that do not exist. This forms a 'sparse' jacobian. Substituting numbers from a subset mechanisms containing the methane and ethane precursors, we get Equation 4.11.

$$\mathbf{J}_{i,j} = \begin{bmatrix} \frac{\partial f_{[C_2H_6]}}{\partial t\ \partial[C_2H_6]} & -2\times10^{-7} & 2\times10^{-7} \\ -0.1 & \frac{\partial f_{[OH]}}{\partial t\ \partial[OH]} & 0.1 \\ & & \frac{\partial f_{[C_2H_5O_2]}}{\partial t\ \partial[C_2H_5O_2]} \end{bmatrix}_{i,j=1}^{3,3} \tag{4.11}$$

This allows us to see two things. Firstly that with the absence of external intervention (e.g. emissions) the overall change of concentration is a conserved property. Secondly ...

Representing these relationships as a simple 'ball and link' style graph gives us Figure 4.13.

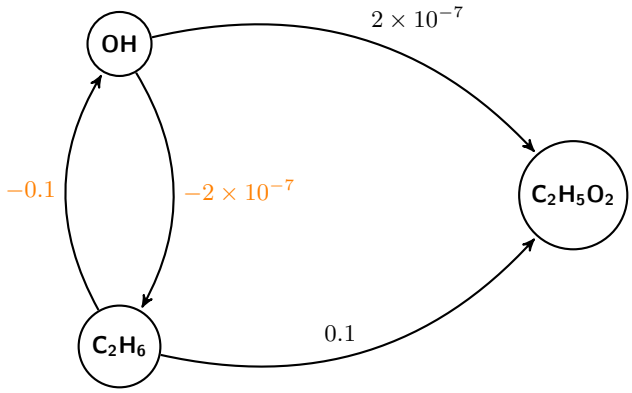


Figure 4.13: A graphical representation of Equation 4.11 derrived from the Equation 4.8

Converting the Jacobian into an adjacency matrix

Adjacency matrixes are a set of matrix representations which can be used in the construction of a graph. The relational data of the Jacobian matrix Equation 4.11 inherently holds such property and can be directly translated to produce a graph, Figure 4.13. However, we notice that some edge weights are negative, which although providing information about the chemical system provides no physical meaning in the graph structure.

It is for this reason that we can reverse the direction for all negative links to produce Figure 4.14.

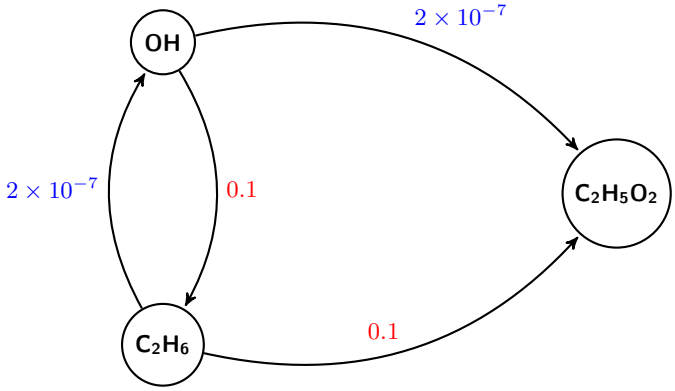


Figure 4.14: Reversing the directions on negatively weighted edges from Figure 4.13

For most graph algorithms this should be sufficient and is generally all that is needed. In some cases, it may, however, be noted that the graph may further be simplified to produce Figure 4.15. Although this is more practical, eigenvector metrics such as PageRank will automatically transfer the ‘flow‘ of information down the system producing much the same overall result.

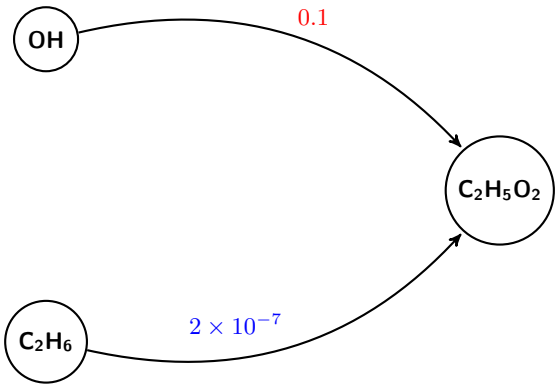


Figure 4.15: Simplifying Figure 4.14

4.5 Case study Example

LONDON ITALY BORNEO BEIJING?

4.5.1 source location using reverse (adjoint) personalised page rank

Bibliography

[lap,] LAPACK — Linear Algebra PACKage. `http://www.netlib.org/lapack/`. http://www.netlib.org/lapack/.

[clo,] Social Network Analysis: Theory and Applications.

[Bastian et al., 2009] Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[Bonacich, 1987] Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182. http://www.jstor.org/stable/2780000.

[Bonacich, 2007] Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social Networks*, 29(4):555 – 564.

[Borgatti, 2005] Borgatti, S. P. (2005). Centrality and network flow. *Social networks*, 27(1):55–71.

[Boudin,] Boudin, F. A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction.

[Brandes, 2001] Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25:163–177.

[Edsu and Ellis,] Edsu and Ellis, D. etudier. https://github.com/wolfiex/etudier.

[Fantin et al., 2012] Fantin, V., Buttol, P., Pergreffi, R., and Masoni, P. (2012). Life cycle assessment of Italian high quality milk production. A comparison with an EPD study. *Journal of cleaner production*, 28:150–159. http://www.sciencedirect.com/science/article/pii/S095965261100388X.

[Freeman, 1977] Freeman, L. (1977). A set of measures of centrality based on betweenness. 40:35–41.

[Freeman et al., 1991] Freeman, L., Borgatti, S., and White, D. (1991). Centrality in valued graphs: A measure of betweenness based on network flow. *Social Networks*, 13:141–154.

[Freeman, 1978] Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239.

[Gemma,] Gemma, J. The Most Influential Men and Women on Twitter 2017. https://www.brandwatch.com/blog/react-influential-men-and-women-2017/. Accessed: 2019-4-28.

[Goh et al., 2001] Goh, K. I., Kahng, B., and Kim, D. (2001). Universal behavior of load distribution in scale-free networks. *Physical review letters*, 87(27 Pt 1):278701.

[Hagberg et al., 2008] Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T., and Millman, J., editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

[Hobson et al., 2018] Hobson, E. A., Mønster, D., and DeDeo, S. (2018). Strategic heuristics underlie animal dominance hierarchies and provide evidence of group-level social knowledge.

[Humphries and Gurney, 2008] Humphries, M. D. and Gurney, K. (2008). Network 'small-world-ness': a quantitative method for determining canonical network equivalence. *PloS one*, 3(4):e0002051.

[Jacob and Lapkin, 2018] Jacob, P.-M. and Lapkin, A. (2018). Statistics of the network of organic chemistry. *React. Chem. Eng.*, 3:102–118. http://dx.doi.org/10.1039/C7RE00129K.

[Jeanningros et al., 2010] Jeanningros, Y., Vlaeminck, S. E., Kaldate, A., Verstraete, W., and Graveleau, L. (2010). Fast start-up of a pilot-scale deammonification sequencing batch reactor from an activated sludge inoculum. *Water science and technology: a journal of the International Association on Water Pollution Research*, 61(6):1393–1400. http://dx.doi.org/10.2166/wst.2010.019.

[Jones et al., 01] Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python. http://www.scipy.org/.

[Kleinberg, 1999] Kleinberg, J. M. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632.

[Krebs, 2002] Krebs, V. E. (2002). Mapping networks of terrorist cells. *Connections*, 24(3):43–52.

[Kumar and Upfal,] Kumar, R. and Upfal, E. The Web as a graph.

[Langville and Meyer, 2005] Langville, A. and Meyer, C. (2005). A Survey of Eigenvector Methods for Web Information Retrieval. *SIAM Review*, 47(1):135–161.

[Mohanty et al., 2014] Mohanty, J. G., Nagababu, E., and Rifkind, J. M. (2014). Red blood cell oxidative stress impairs oxygen delivery and induces red blood cell aging. *Frontiers in physiology*, 5:84. http://dx.doi.org/10.3389/fphys.2014.00084.

[Needham and Hodler, 2019] Needham, M. and Hodler, A. E. (2019). Practical Examples in Apache Spark & Neo4j. *O'Reilly*.

[Oliphant, 2006] Oliphant, T. (2006). Guide to numpy.

[Ottens et al., 2006] Ottens, A. K., Kobeissy, F. H., Golden, E. C., Zhang, Z., Haskins, W. E., Chen, S.-S., Hayes, R. L., Wang, K. K. W., and Denslow, N. D. (2006). Neuroproteomics in neurotrauma. *Mass spectrometry reviews*, 25(3):380–408. http://dx.doi.org/10.1002/mas.20073.

[Page et al., 1999] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. *Stanford InfoLab*, (1999-66). Previous number = SIDL-WP-1999-0120 `http://ilpubs.stanford.edu:8090/422/`.

[Pokroy et al., 2009] Pokroy, B., Epstein, A. K., Persson-Gulda, M. C. M., and Aizenberg, J. (2009). Fabrication of Bioinspired Actuated Nanostructures with Arbitrary Geometry and Stiffness. *Advanced materials*, 21(4):463–469. http://doi.wiley.com/10.1002/adma.200801432.

[R. Seeley, 1949] R. Seeley, J. (1949). The net of reciprocal influence: A problem in treating sociometric data. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 3:234–240.

[Sabidussi, 1966] Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4):581–603.

[Spencer et al., 2010] Spencer, R. G. M., Hernes, P. J., Ruf, R., Baker, A., Dyda, R. Y., Stubbins, A., and Six, J. (2010). Temporal controls on dissolved organic matter and lignin biogeochemistry in a pristine tropical river, Democratic Republic of Congo. *Journal of geophysical research*, 115(G3):2069. http://doi.wiley.com/10.1029/2009JG001180.

[Vigna, 2016] Vigna, S. (2016). Spectral ranking. *Network Science*, 4(4):433–445.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

[Wood, 2014] Wood, B. (2014). The Origin of Humans Is Surprisingly Complicated. *Scientific American*. `https://www.scientificamerican.com/article/the-origin-of-humans-is-surprisingly-complicated/`.

Contents