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*

Contents

4.1 Introduction

The node-link (ball-stick) [REF SECTION] 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 [citep]. It is for this reason that the node-link visualisation format has been used for anything ranging from transportation maps [citep BECK] to the differentiation of ancestorial lineages of the human race (Figure 4.1). However, the abundance and complexity of real-world data often present us with difficulties in manually representing it in a useful form. In SECTION XX it is suggested this may be overcome with the use of computational analysis and automated visualisation tools. Such methods usually require a level of data manipulation to transform the data into a machine parseable form.

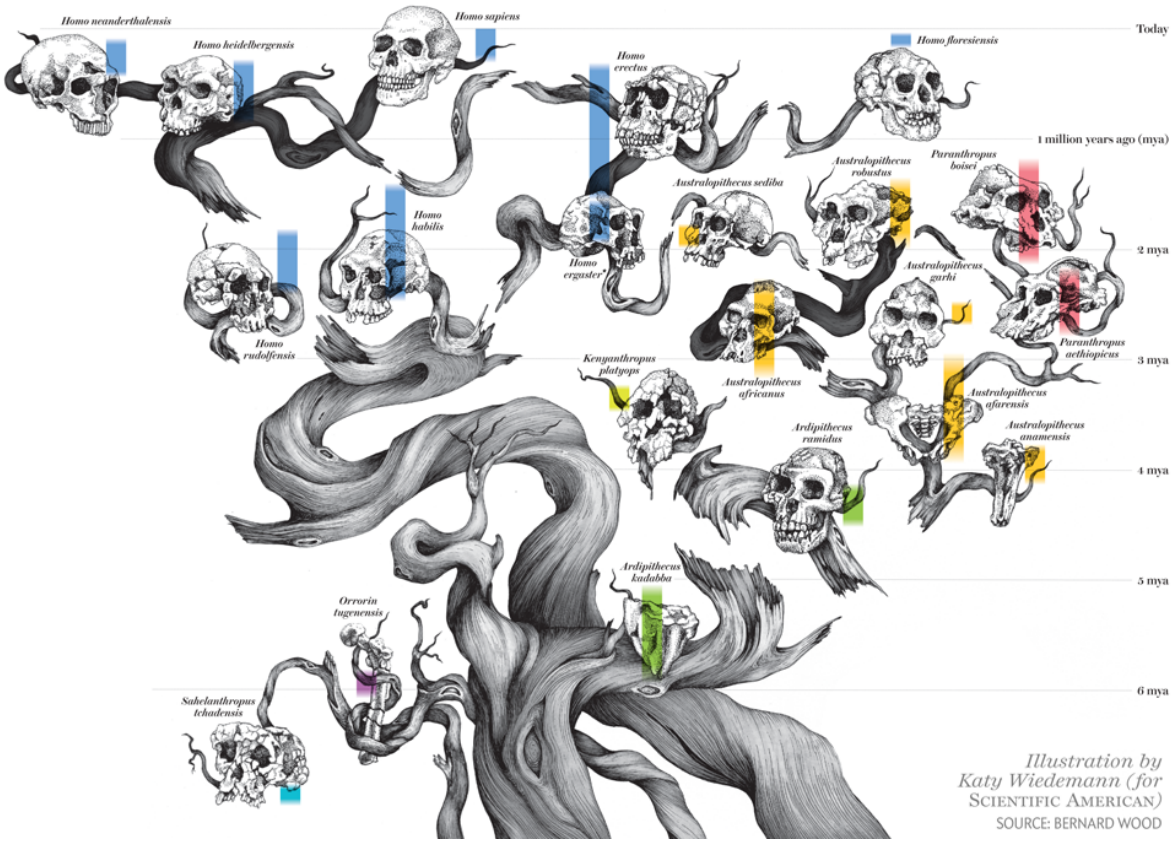


Figure 4.1: **The human family tree.** This is a visual depiction of the human lineage, starting with our common ancestorial roots. In SECTION it was shown that the use of trees / graphs[1]is useful in showing relationships between items. Source: [Wood, 2014]

In the field of mathematics a graph, $G(\nu, \epsilon, \omega)$, is defined as a function of items (vertices[2]), ν which are connected through a series of connections (or edges[1]) representing any relationships between them, ϵ.

[1] A tree is a special case of a graph

[2] 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

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

This chapter builds on the work shown in SECTION XXX - where the ability to represent complex data in the form of a graph was used to (visually) draw information regarding network structure and temporal changes. Here I will begin by exploring situations where the visual representation of many, large or complex networks is impractical. We start by introducing a series of mathematical approaches which are capable of quantifying the graph (and nodes within it) and apply them to the co-author network for papers regarding the Master Chemical Mechanism, section 4.2. Following these global metrics are used to categorise the chemistry within different mechanism subsets, and provide us with an insight to the chemistry structure (SECT LABEL) and finally apply these to real-world simulations representing a range of environments (marine, rainforest and urban) in SECTREF.

This allows for a higher level of automated analysis which can be used to batch process, analyse and categorise chemical simulations. section 4.2 begins by introducing the most common of the graph metrics which can be used for analysis. To do this a citation graph is generated by web-scraping google scholar results.

4.2 Graph Metrics

An increase in the ability to gather and store data results in a difficulty to understand it (ref SECTION). The production of large, multivariate networks of inexplicable complexity greatly hinders our ability to draw out meaningful conclusions based on visualisation alone. This means that much like the generation of mechanism, or creating semi-automated graph drawing layouts, we must rely on the field of mathematics coupled with computational aid (REF SECTION).

Numerical algorithm, derived from the field of Graph Theory can be used to circumvent the need for individual graph analysis and provide us with information about the network. One such subset of numerical algorithms are regarded as "centrality metrics", and may be used to rank the role and importance (centrality) of a node. In the following sub-section, the most common (REF PAPER) centrality metrics are discussed and applied to the MCM citation network.

4.2.1 Centrality metrics and academic publishing.

One common application for graph analysis and visualisation is the representation and prediction of citation counts within academic journals [Small, 1973; Page et al., 1999; Monastersky and Van Noor-

den, 2019; Molontay and Nagy, 2020]. Here network-visualisation techniques may be used to highlight the origins of a paper - for instance, Figure 4.2 shows the multi-disciplinary research which underpins 6 prominent discoveries in the last 150 years.

To the properties presented by different centrality metrics (described above), we apply them to an approximate representation of the citation graph relating to the Master Chemical Mechanism (subsection 4.2.2).

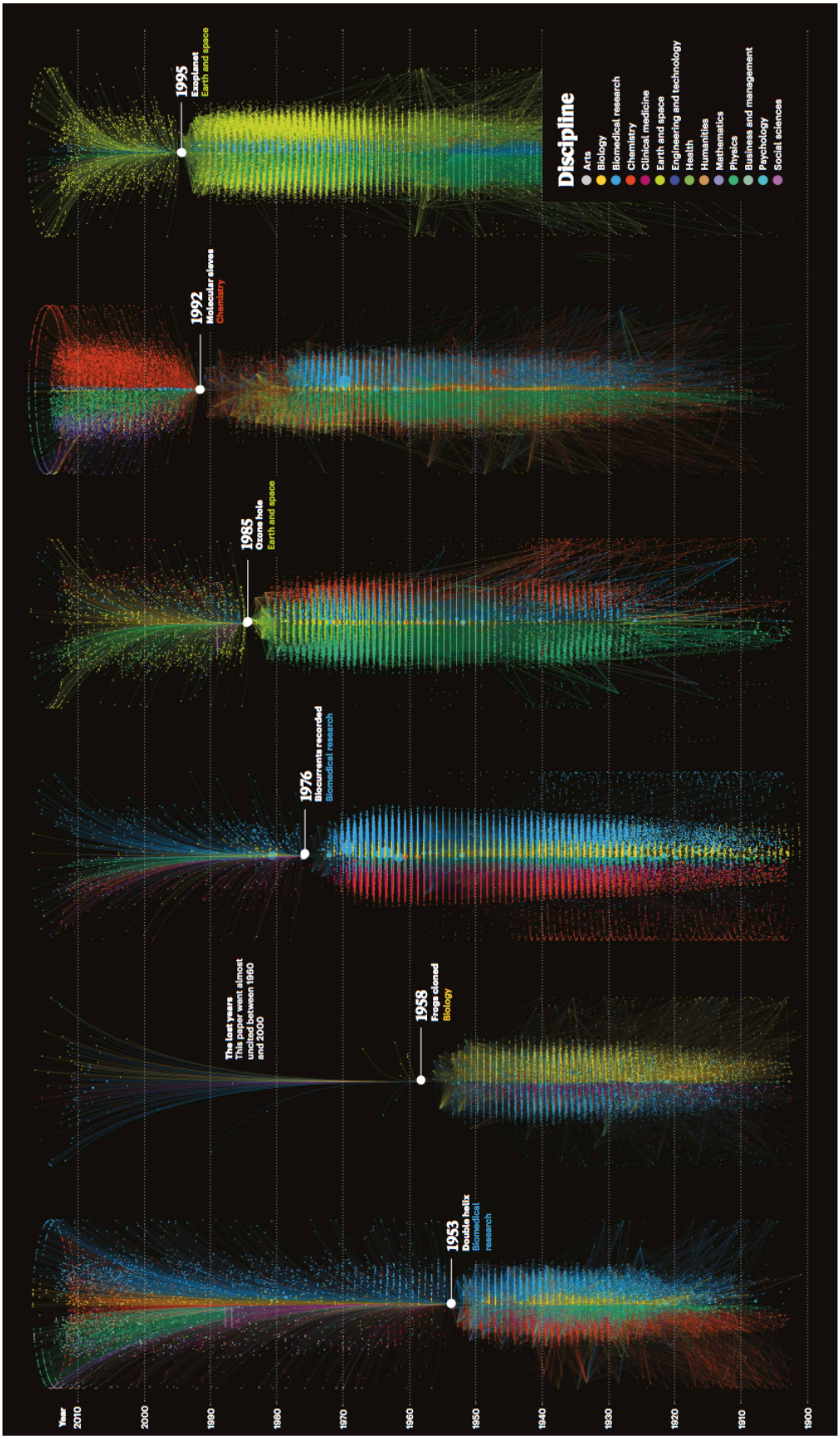


Figure 4.2: **150 years of letters to Nature.** A visualisation showing how previous research is used to inspire future studies. Important discoveries (DNA, Cloning(frogs), Bio-Currents, Ozone Hole, Molecular Sieves and Exoplanets) are split into research which contributed to their formation (below), and the consequent papers produced from each discovery. Use of colour is used to emphasise the multi-disciplinary nature of prolific scientific discovery. Source: [Barabási, 2019]

4.2.2 The Master Chemical Mechanism (MCM)

The MCM, [**?**], is a near explicit representation of our foremost understanding of gas-phase tropospheric chemistry. The mechanism describes the oxidation of 143 primary emitted VOCs and the respective rates at which this occurs. It has been used in the...

Information on the chemistry, - x species - y ... first published and how this can be used with regards to the following algorithms are presented in REF JENKINS 15 ACP.

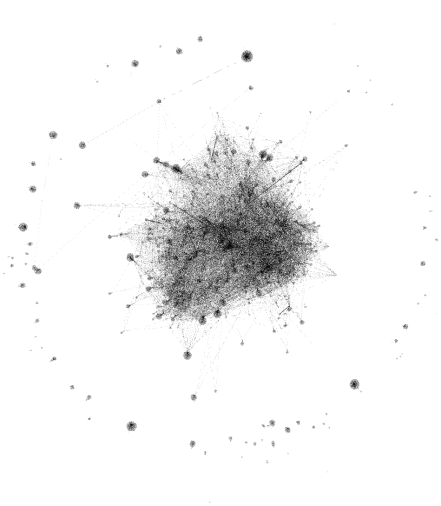
4.2.3 Data Collection

To generate a dataset on papers related to the MCM. The academic search engine (Google Scholar [Google, 2019]) is queried for all articles containing the words { *"Master", "Chemical", "Mechanism"* and *"MCM"* }. For each match, the first 100 pages of results are selected. Each of these contains 10 articles, from which the first 100 pages of related articles are chosen. In taking the top 1000 citations for each page a network of 15744 papers and 30178 citations[3] is created. This process made use of an edited version of the *etudier* Github repository, [Edsu and Ellis, 2019].

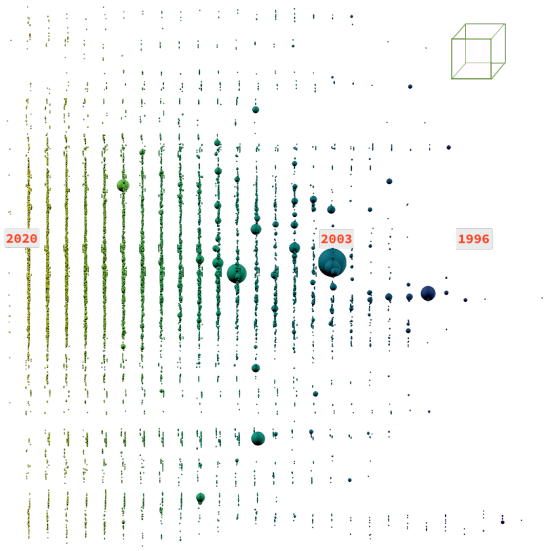
4.2.4 Visualising the data.

The initial visualisation of the dataset is accomplished through the use of THREE.js [Cabello, 2019]. This makes use of WebGL bindings and allows for the efficient viewing, querying and interacting of the data in 3 dimensions. This helped identify the temporal changes within the network by mapping a papers publication year to the z direction, Figure 4.3, as discussed in subsection 4.2.5.

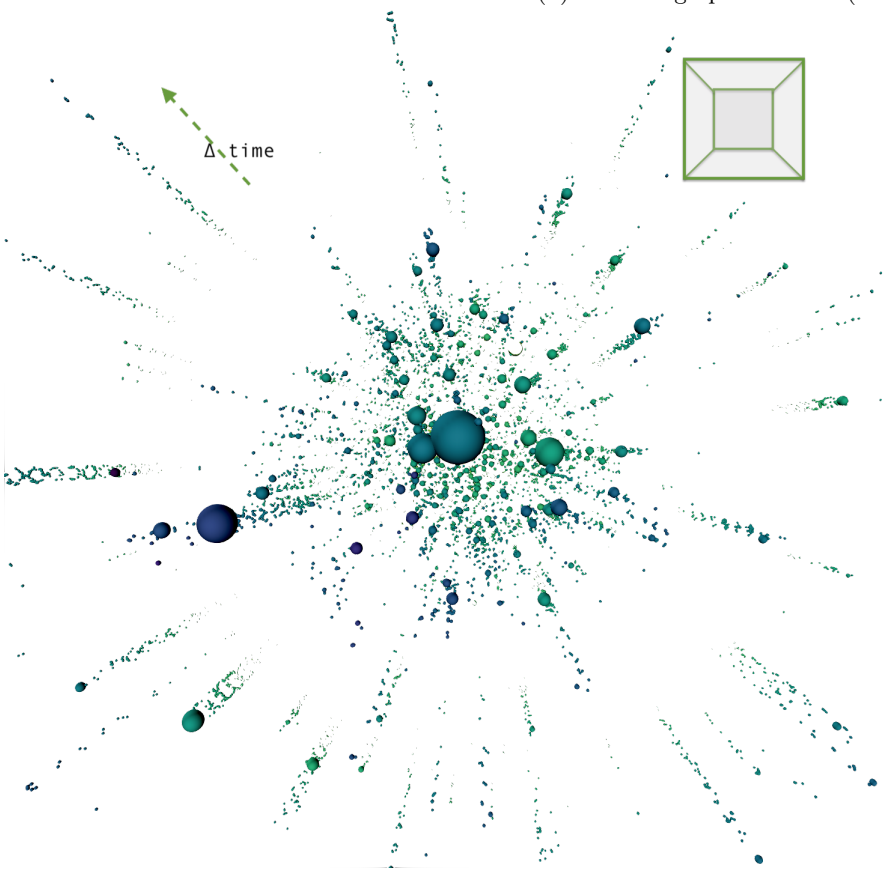
[3]Note: this had the potential of returning up to 1000,000 nodes



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



(b) 3D orthographic camera (sideview)



(c) 3D perspective camera

Figure 4.3: **Initial 3D graph representation of the scraped MCM citation graph.** (a) shows the 'classic' graph representation of the network. (b) shows a size representation using an orthographic perspective. Here time is shown across the x axis, with yellow being the most recent. (c) uses a perspective camera, which emphasizes the... 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.

4.2.5 Filtering the data

In the method used to web scrape data, there are several features which need to be corrected/removed. The reasons for this are discussed below.

Pre-1996

There exist several papers predating the conception of the MCM (1996). A number of these can be attributed as incorrect data, with publication dates <1900 which may be the result of missing information or a fault in googles web scraping algorithm. Any such papers are removed from the dataset.

For otherwise correct articles, those published pre-1996 are also filtered from the dataset - this is because we are interested in identifying the influence the MCM has had on research and not the research that may have led to its creation. This can be seen in the cone-like shape emanating from the first MCM papers in Figure 4.3b.

N-th degree research

Not all research articles in a field reference other articles with the same field. Figure 4.2 showed us that many of the great discoveries in science have a multidisciplinary nature. It is for this reason that it is expected that articles from non-atmospheric areas of research may reference or build upon specific areas of research touched by the MCM. Such papers, and in consequence the papers which cite them, have little or no links to many of the core MCM papers. Such papers manifest themselves as a halo of satellite clusters which are connected by themselves but not with the main body of the graph, Figure 4.3a. In using a 3D perspective viewpoint (Figure 4.3c) it is possible to identify the paper which references the MCM and then the consequent papers which cite it by observing the satellite clusters, and the gradually lightening spiral of papers which emanate out of it.

Analysis of the network connections for each cluster can allow us to identify the indirect relationships some of these diverse topics (Table **??**) contained within the satellite nodes. Here it can be seen that the use of photochemical ozone creation potentials [Derwent et al., 1998; Jenkin and Hayman, 1999] are used for the Life cycle assessment of Italian high-quality milk production [Fantin et al., 2012]. Similarly indirect paths such as the paper: "Temporal controls on dissolved organic matter and lignin biogeochemistry in a pristine tropical river" ([Spencer et al., 2010]) can be used to link to [Stubbins et al., 2008] and ultimately the MCM protocol paper [Saunders et al., 2003].

If we desired to remove such papers, the simplest method would be to recreate the graph into one

where links are drawn between papers that are cited together (subsection 4.2.6) and then removing any nodes without any external connections (isolates).

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: A selection of research papers not directly connected to the field of atmospheric modelling.

Unprobable occurances

Finally, the extracted network also contains many disconnected component subgraphs - graphs with no connection to atmospheric science. An example of this is seen in an article about neuroproteomics in neurotrauma [Ottens et al., 2006]. In analysing the paths which connect this, it is seen to cite the paper on "Large scale gene expression profiling of metabolic shift of mammalian cells in culture", [Korke et al., 2004]. This is an anomaly which within its structure contains the words "Master", "Chemical" and "Mechanism" (separately) and has ‘MCM’ as an abbreviation for one of the author names. To remove such papers, all disconnected sub-components are removed from the analysis.

A note on unintentional filtering

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.

4.2.6 The Co-citation Network

The document coupling techniques of co-citation was introduced in the 1970s as an alternative approach for quantifying the results within the science citation index [Small, 1973]. Rather than rep-

resenting a graph using backpropagation (through the use of referencing and citation counts), a co-citation network introduces a link between papers if, and only if, they have been cited together. Although this loses the directionality of a graph, it allows us to show forward propagating trends between papers within the same field.

Applying the above method allows us to reduce the citation graph of 451 papers and 5402 edges to an undirected co-citation graph of 2758 edges - halving the number of original links between papers.

4.2.7 The Co-authorship network

An alternative to exploring which papers which are cited together are to look at their authors. Here undirected links are drawn between authors on the same paper. This style of analysis was used to show that the number of papers per author, and the total number of authors per paper can vary between research fields, [Newman, 2004]. In combining this with a series of network centrality metrics, [Fujita et al., 2017] revealed that it is possible to discern promising researchers from both iter and Intra disciplinary groups.

In building a co-authorship network for the MCM, we can identify authors who publish together[4] and highlight research groups who work with the MCM, **??**. This shows how authors with a similar geographic location/institution are more likely to publish together. The largest cluster here falls under the MCM developer team, which resides between the Leeds and York universities. Next two German institutions which are heavily involved in the atmospheric chemistry field (Julrich and Max Planck), followed by an assortment of Chinese authors, mainly centred around the Beijing or Hong Kong region.

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

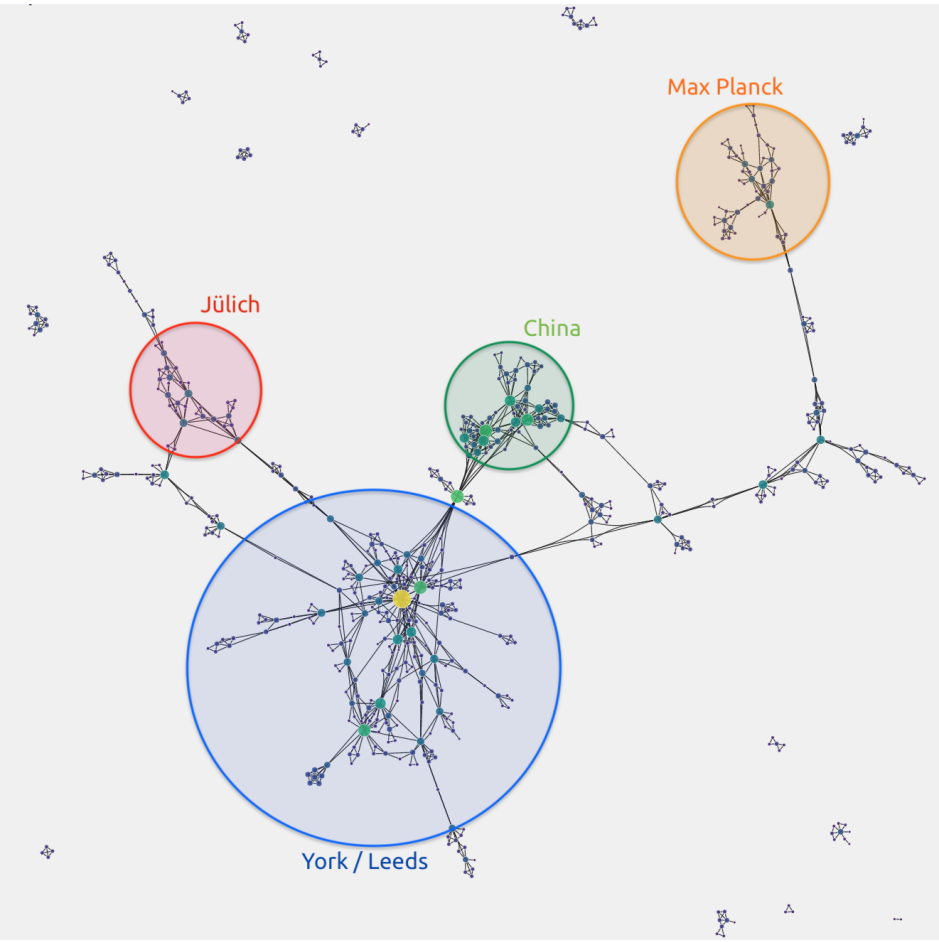


Figure 4.4: **The co-author network.** In representing the authorship network as a force directed graph we are able to see cliques or clusters of people who publish together. It can be noted that this often occurs when they have a similar geographical location.

4.3 Metric analysis

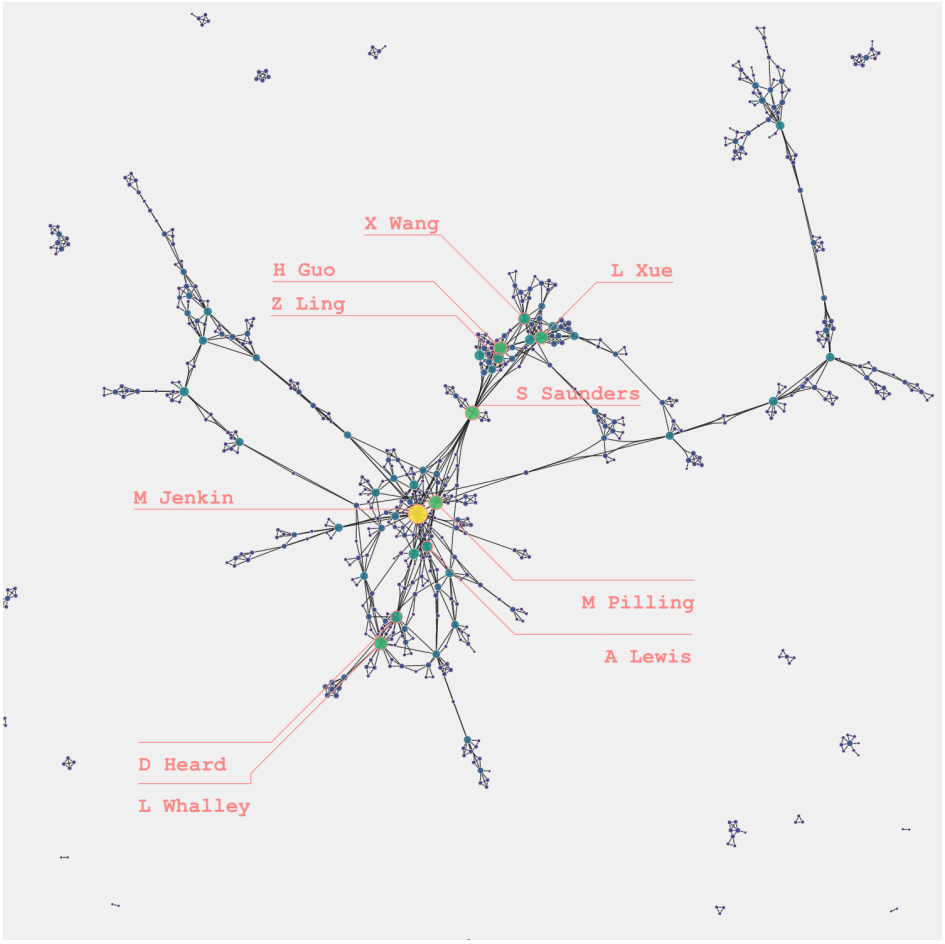
To demonstrate the information provided by different centrality metrics, a simple intuitive network (the co-author network in **??**) shall be used. This subsection will access the efficiency of graph centrality metrics in their ability to identify important nodes within a network.

4.3.1 Degree Centrality

The simplest, and most intuitive, metric is degree centrality [Freeman, 1978]. This is described as the sum of all links incident on a node - simply put, we count the number of edges going in and out of a node. This gives us an idea of the importance of a node and has been used to calculate influence within social media or the probability of a profile committing online auction fraud [Gemma, 2019; Freeman, 1978].

For the author network, Figure 4.5 we see that many of the names on the list are either contributors

to the MCM or have worked with them at Leeds. It is also seen that the authors with the most collaborations, or links, are very likely to appear within the most cited or citing papers (Table 4.2 and Table 4.3 discussed below). This is likely because both development (well-cited) and the evaluation/usage (well citing) of a mechanism requires knowledge from a range of different fields, making it an interactively collaborative process.



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

Figure 4.5: **Degree Centrality.** In applying the degree centrality to the co-authorship network, it is possible to pick the authors with the greatest number of papers, of which the top 10 have been listed.

Directed Degree

For graphs where link direction holds an inherent meaning regarding their representation (for example in the citation graph an outward link symbolises that paper citing the one that the link points to), it is possible to further divide the degree centrality metric into inwards and outward links. This can allow us to separate items which provide a large number of lots of information (in-degree) and those who collate or collect it (out-degree). In applying these metrics to the directed citation graph, it is possible to get an insight into the core MCM development papers (Table 4.2) and separtate them from those which make use of the mechanism as part of a greater study (Table 4.3).

Protocol for the development of the Master Chemical Mechanism, MCM v3 Part A tropospheric degradation of nonaromatic volatile organic compounds	Saunders et al. [2003]
Protocol for the development of the Master Chemical Mechanism, MCM v3 Part B tropospheric degradation of aromatic volatile organic compounds	Jenkin et al. [2003]
Development of a detailed chemical mechanism MCMv3. 1 for the atmospheric oxidation of aromatic hydrocarbons	Bloss et al. [2005]

Table 4.2: **In-Degree of the citation network**: The top 3 most cited papers.

The MCM v3.3.1 degradation scheme for isoprene	Jenkin et al. [2015]
Atmospheric photochemical reactivity and ozone production at two sites in Hong Kong Application of a master chemical mechanismphotochemical box model	Ling et al. [2014]
HOx budgets during HOxComp A case study of HOx chemistry under NOxlimited conditions	Elshorbany et al. [2012]

Table 4.3: **Out-Degree of the citation network**: The top 3 most citing papers.

4.3.2 Closness Centrality

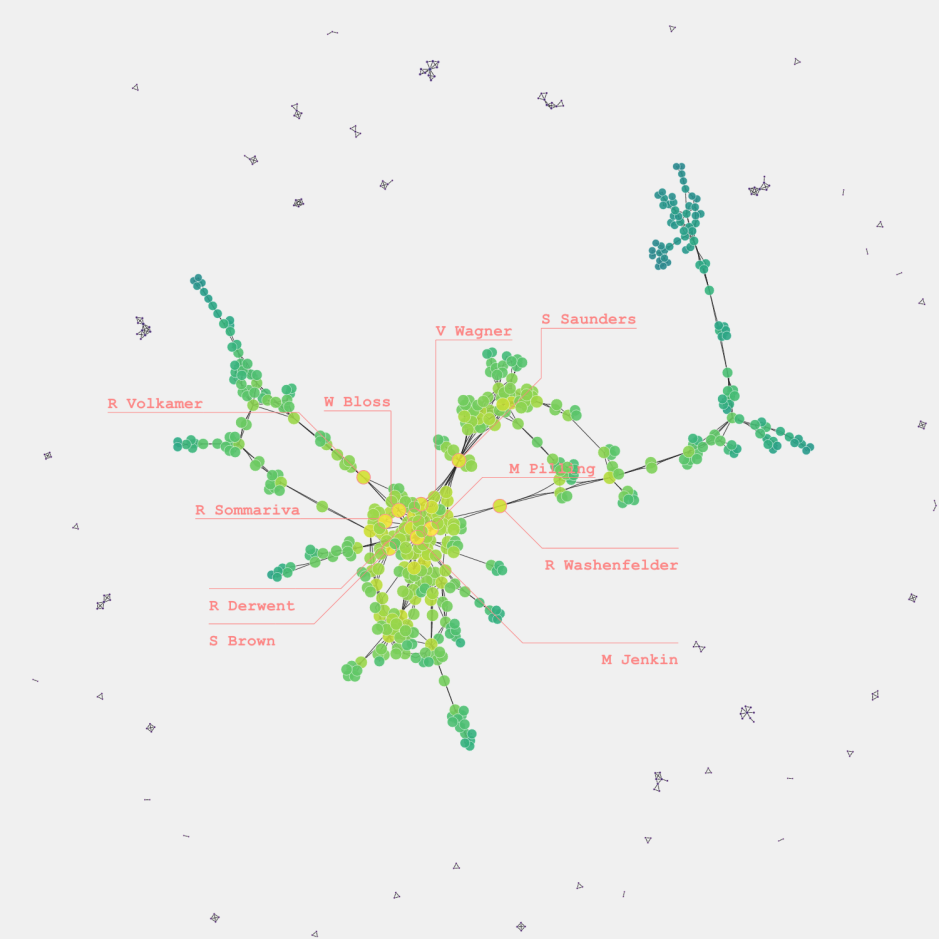
Often within a network, we are interested in how easy it is to to get information from one node to every other node. This is what the closeness centrality tells us. To calculate a nodes closeness we begin by taking the reciprocal sum of all the Dijkstra paths[5] to every other node [poliaktiv, 2011; Sabidussi,

[5]The shortest available path.

1966]. This gives is a representation of how far information from a certain will need to travel to reach every other node. Such a metric has applications in intelligence gathering, telecommunications and word importance within key-phrase extraction [Krebs, 2002; Borgatti, 2005; Boudin, 2013].

> ***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.*

For the co-authorship network, Figure 4.6, nodes have been coloured by their closeness value. Here a heat-map-like effect may be observed, showing that information between the dense Leeds-York cluster is easier to disseminate across all parts of the graph than that of localised branches of authors less involved with the development team. The results of the closeness centrality suggest that should a problem (bug) or improvement (update) occur, Michael Pilling would be the best served to pass that information to all other groups using the MCM.



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

Figure 4.6: **Closeness centrality within the co-Author network.** Here a colour/size gradient is seen, with the nodes that are more central (in location) and better connected having a higher closeness than those in the peripheries - which are harder to get to.

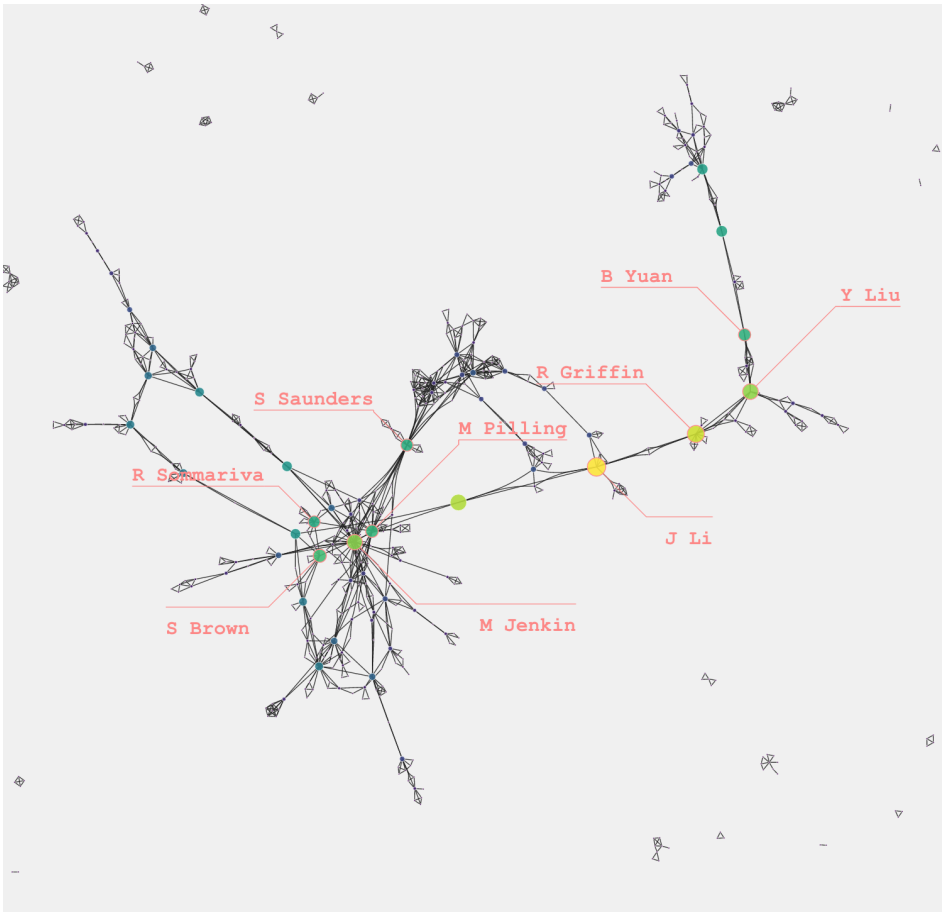
4.3.3 Betweenness

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. Nodes with a high betweenness control, or limit, the amount of information that can be transferred across the network. If a node lies on a geodesic (the shortest path between two other nodes), we may consider it a 'pivotal' node, due to its role within the network [Needham and Hodler, 2019]. Should such a node then be removed, the

overall flow of information incurs either a deviation, the information will either need to travel a longer (alternative) route or may not be able to reach its destination at all [Freeman et al., 1991; Freeman, 1977; Brandes, 2001; Borgatti, 2005]. Betweenness centrality is a count of the number of geodesics which pass through a node. If multiple 'shortest' paths are possible, this is accounted for within the denominator.

> ***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.*

Authors with a high betweenness in Figure 4.7 are seen to lie along the joints between clusters. Here we can imagine that removing Li, Griffin or Liu can disrupt the overall flow of collaboration, potentially isolating the work of the Max Planck from that of everyone else. Similarly, Jenkin and Pilling can be seen as holding much of the Leeds cluster together. In removing them from the network (if for example, the refused to collaborate) it is possible to see how many of groups within the Leeds environment may not have worked together, with the cluster potentially separating into several smaller groups. Finally, we see Saunders (Australia), who served to introduce the MCM to the Chinese atmospheric community. In removing her from the network, it can be seen that much of the collaboration which exists would have been significantly less likely.



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

Figure 4.7: **Betweenness centrality within the co-Author network.** Nodes which lie on a pivotal position (connecting/bottleneck) tend to have a high betweenness value due to their crutial role within the network.

4.3.4 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 Table 4.3.4, 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 Normalisati
No Damping	Eigenvector [Bonacich, 1987, 2007]	Markov Chain State [R. Seeley,
Damping	Katz [Goh et al., 2001]	Total Effect PageRank [Pag 1999]

Here damping terms represent the probability of moving to the new random starting position, allowing for the user to 'randomly select a new webpage' or leave an isolated cluster. The 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 is discussed in more detail in the next subsection.

Hypertext Induced Topic Search (HITS)

A common eigenvector algorithm used for classifying webpages is the HITS algorithm. This helps categorise the role of a node as either a Hub or an Authority, [Kleinberg, 1999; Langville and Meyer, 2005; Kumar and Upfal, 2000]. Similar to the in and out-degree metrics, this algorithm separates nodes with many outgoing links (an authority) from those with many ingoing ones (an information hub). Overall this provides similar results to the in/out degree, although since it looks more on how information propagates across the network as a whole, it often provides more accurate, and different, rankings to simple degree analysis.

4.3.5 Page Rank

Arguably the best-known centrality algorithm is PageRank. This is a spectral method for measuring the transitive influence of a node, by taking the effect of neighbours and by their neighbours into account [Needham and Hodler, 2019]. The page rank algorithm was initially developed to provide a better way of ranking web pages [Page et al., 1999]- here an important page is not only one of many links, but links to other important sources. In the context of academic papers, that same paper also found that in predicting future citations, the page rank algorithm fared better than using the current citation count of a paper. To explain how this works, we will look at the mathematics behind the algorithm, and then eventually apply it to the co-authorship graph in 4.3.5

The Google Matrix

To solve for page rank, a google matrix must first be constructed. Once done this is iterated until convergence is reached.

To build a google matrix, we must first generate a dyadic link map of the graph[6] - its adjacency matrix $A_{i,j}$ (i, j are the source target indexes). This is then converted into a Markov matrix $M_{i.j}$ by dividing each column j by the sum of the total outgoing links of node j, Algorithm 1. Species with no outgoing links (sinks), are adjusted with either a personalised list of values or the constant $1/n$, (where n is the number of nodes) to replace the zero-sum columns. This produces a normalised[7] matrix of Markov chains representing the fractional production for node j from all other nodes.

Algorithm 1 Adjacency to Markov matrix.

1: Obtain graph adjacency matrix, $A_{i,j}$.
2: **repeat**
3: **for each** $j \in$ columns **do**
4: $M(:, j) \leftarrow A(:, j)/\Sigma_{i=1,n}A(j, i)$
5: **end for**
6: **until** $\Sigma_{i=1,n}M(i, j) = 1$

The google matrix $G_{i,j}$ 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. To account for this, a damping factor, typically $\beta = 0.85$, is used. This defines the probability that the user follows a link, and that for which they randomly select another page: $(1 - \beta)$ [8]. The damping factor used varies greatly with the application, with values such as $\beta = 0.694$ having been found optimal for the use of biological data [Hobson et al., 2018].

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

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*

[6] In sociology a dyad is a group of two people - the smallest possible social group.

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

[8] Also known as teleportation.

Solving the algebra

Once defined, the google matrix is solved by propagating a one's vector, r of length n, where n is the number of species using Algorithm 2.

Algorithm 2 Solving the google matrix linear algebra

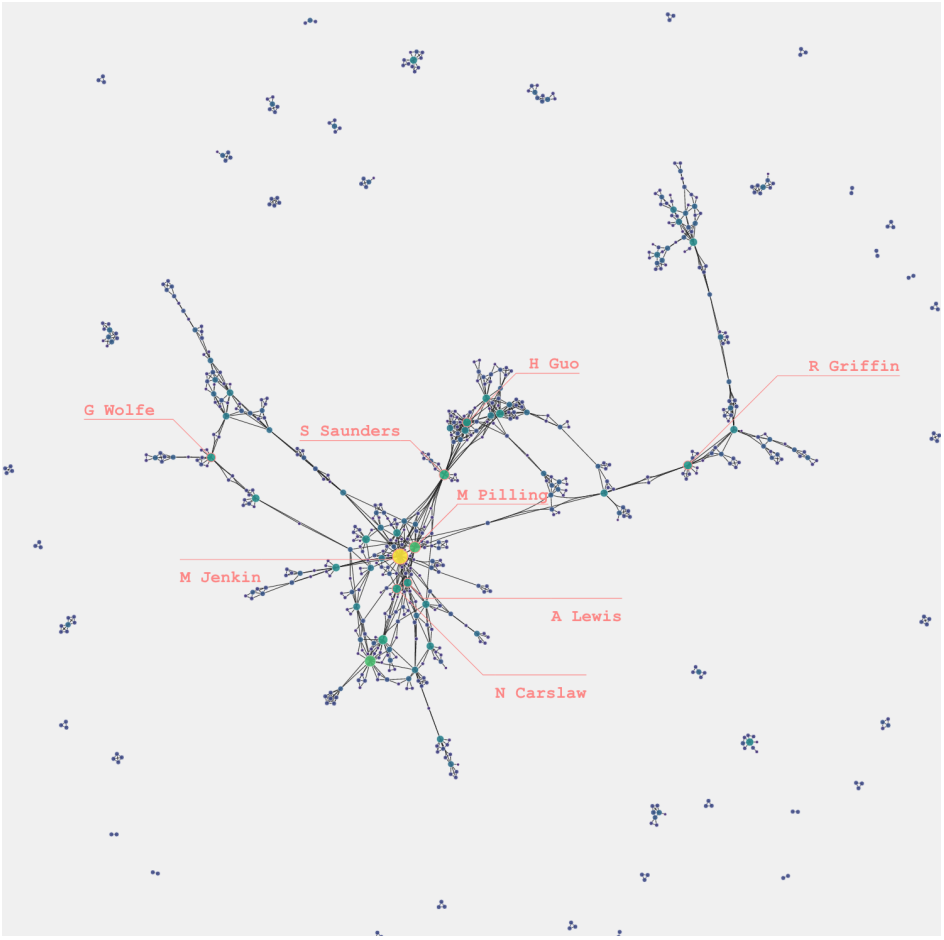
1: Define value vectors $\bar{r}_t$ and $\bar{r}_{t+1}$:
2: $\bar{r}_t = [1_1, 1_2, ..., 1_n]$, $\bar{r}_{t+1} = [0_1, 0_2, ..., 0_n]$
3:
4: **while** $||\bar{r}_{t+1} - \bar{r}_t|| > \epsilon$ **do**
5: $\quad \bar{r}_{t+1} \leftarrow M.\bar{r}_t$
6: $\quad \bar{r}_t = \bar{r}_{t+1}$
7: **end while**

This is repeated until a pre-defined tolerance, ϵ is reached. For best results, this can be set to just under the numerical precision of the programming language/hardware.

For smaller systems, it is possible to use the LAPACK [lap, 2019] library, as used by [Oliphant, 2006]. For a large network, however, the computation of an $n \times n$ matrix can be very memory inefficient for small machines. It is then possible to apply the methods as described above using a sparse matrix on per-node bases as can be seen within the scipy implementation of the networkx source code [Jones et al., 01 ; Hagberg et al., 2008].

Prediction

As the PageRank algorithm is a physical representation looking at how quantities 'flow' within a network, it can be used to identify not only the bottlenecks (betweenness centrality) but also any nodes which are connected well within the network. As the flows between a node are somewhat governed by the number of links it contains, the PageRank algorithms tend to correlate, but not a dependance, on the betweenness of a node. Figure 4.8 shows the PageRank algorithm to identify important authors within each 'cluster' or research group. Due to its propagating nature authors connected to these important nodes are often also of greater importance. An application of this can again be the determination of how to best spread new results or information with the least number of people. *Note: if we only had one person we would probably use the node with the highest closeness centrality.*



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

Figure 4.8: **Page Rank centrality within the co-Author network**.

4.3.5.1 Conclusions

In this section, we have explored the use of centrality metrics to provide us with information on an unweighted co-authorship network of the MCM. Having used these to demonstrate the different roles that may be extracted from a node, we can move on to applying them to a chemical mechanism. In the next section, a global set of metrics will be used to determine the network type/structure of the MCM. Once this has been done, graph construction using simulation results (a weighted graph) will be looked into in subsection 4.5.1.

4.4 Classifying the Master Chemical Mechanism network

Having shown that graph metrics can help the roles of individual nodes within the network, I will now apply them to a chemical system. Since computational efficiency and resources are often a limiting factor, many applications of the MCM only require a small subset of the entire mechanism. For this reason, it may be of interest to compare these against each other, in an attempt to classify the type of network the MCM chemistry falls under. In this section, we apply graph theory to the entire MCM network, to determine its defining characteristics. This is achieved through the analysis of several hundred randomly selected subsets of the MCM.

4.4.1 Network density

Network density is the easiest to understand. Visually this can induce complexity and obscure aspects in a graph, mathematically it can greatly increase the computation time for metrics or algorithms. By definition, we can define network density as a measure of how well connected a node is to every other node, mathematically it is the ratio of edges against the total number of possible edges for a complete graph[9] of the same size. In chemical terms, we can use this to determine the sparsity of the graph (which has applications on model integrator selection) and give us insights on the chemical structure. In Figure 4.9 the addition of more species (nodes) results in an overall decrease in the node-edge ratio - it's density. This suggests a modular or hierarchical structure, where new species directly react only with a set number of species, and not the entire mechanism. An explanation for this is that the addition of larger species introduce new branches within the chemistry, which then need to be oxidised before they are small enough to react with the species from a different branch. 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.

[9] A complete graph is one where every node is connected to every other node.

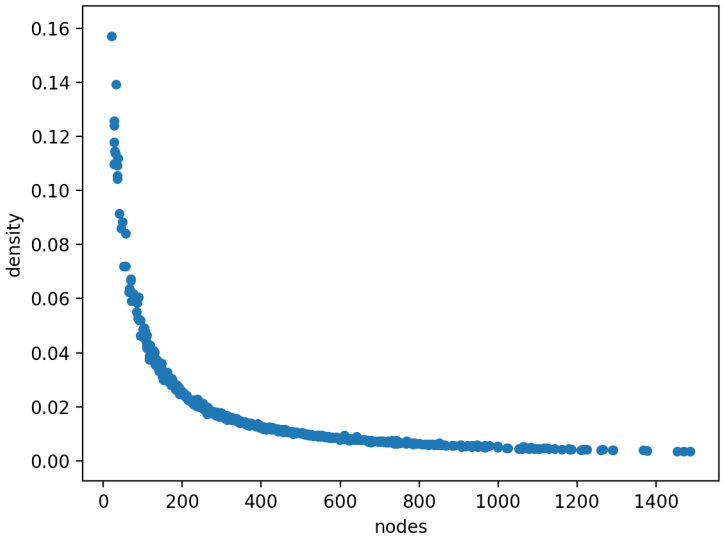


Figure 4.9: **How the MCM graph density scales with number of species.** A figure showing that an creasing the number of species within a mechanism subset results in an increased model sparsity (decreasing density).

4.4.2 Small world Phenomena

Within the biological or social sciences the small world phenomenon, colloquially known as 'six degrees of separation', is a common occurrence within network structure [Watts and Strogatz, 1998]. Such networks have a large number of localised clusters (cliques) all with a short path length between their elements [Humphries and Gurney, 2008]. This makes it easy to reach all parts of a network with only a couple of hops/reactions. In the initial interactive explorations of graph visualisation, it was found that in selecting the reactions of a node, and consequently the reactions of all the nodes which react with them, very quickly a large proportion of the network was highlighted. This suggests that the network may follow the small world phenomena, especially as it is a sparse network, subsection 4.4.1.

One of the possible methods for establishing the small world-ness of a graph falls under the of the omega (ω) coefficient:

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

Here C is the average clustering coefficient and L, the shortest path length of the graph. Comparing these with the average shortest path length, L_R, and clustering coefficient C_l (as calculated using an equivalent random and lattice graph) gives the above equation. The output is a result between positive and negative one {-1,1}, where a value of 0 suggests the graph exhibits perfect small world-ness.

In assessing the network structure of the MCM, a Monte Carlo (random) approach was taken to

extract several hundred subsets from the entire mechanism. For each of these, the omega coefficient was calculated and plotted in Figure 4.10. Here it is seen that subsets with a small number of species (for example those derived only from Methane or Ethane) exhibit a more lattice-style graph, with the majority of the networks showing a more random network structure. All the results, however, show a prevalence of small-world features over any of the alternative network structures - they are closer to 0 than 1 or -1. This reflects the idea that large species react locally, forming branches (REF VIS CHAPTER), before oxidising to smaller species with more reactions. This result is also seen within the Reaxys chemical database [Jacob and Lapkin, 2018].

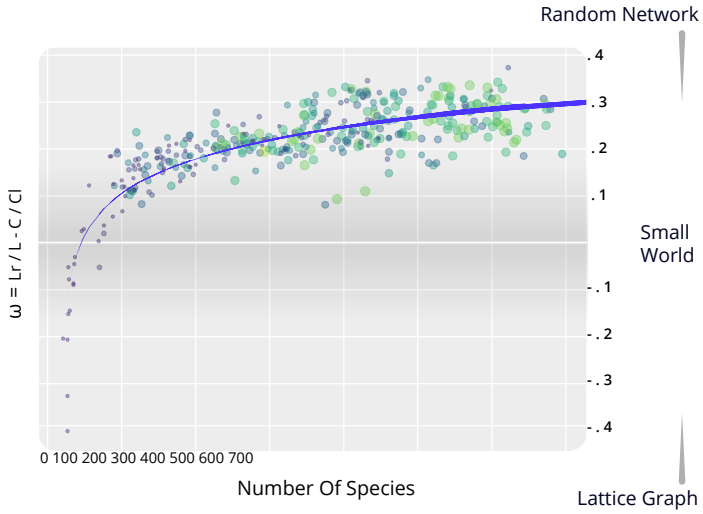


Figure 4.10: **A figure showing the small worldness for many Monte-Carlo selected MCM subsets**. The network structure of these is then assessed using the omega coefficient, with [-1,0,1] corresponding to the perfect lattice, small-world and random network structure. Here Node size and colour represents the number of reactions in the subset and the number of primary VOCs (blue=small, green=large).

4.4.3 Power Law and Scale-free graphs

In real-world applications, it is common to have a hierarchical structure. These are often seen in the increase of citation counts in academic papers [de Solla Price, 1965], email threads [Ebel et al., 2002] and the world wide web [Needham and Hodler, 2019]. Unlike random or small-world graphs, scale-free graphs take a hub-and-spoke structure (Figure 4.11), which follows a power-law distribution - that is that scaling probability $p(x) \propto x^{-\alpha}$, where α is a constant and known as the scaling parameter.

Broido and Clauset [2019] suggests that scale-free networks are rare, and often misdiagnosed with incorrect tests, or the misinterpretation of power-law features in a network. Similarly, Clauset et al. [2009] suggests that even if the data distribution of a graph is well represented by the power-law distribution, in many cases a logarithmic or exponential distribution may have a better fit.

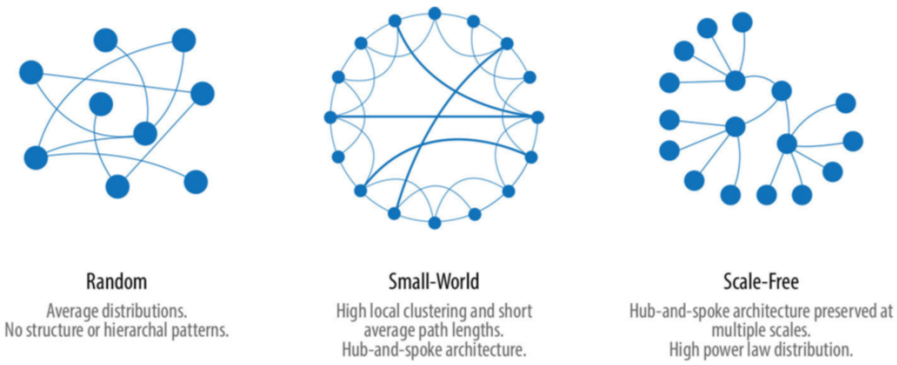


Figure 4.11: **The different network structures**. A visual depiction of the different graph structures. Source: Needham and Hodler [2019]

To assess the best distribution for describing the monte carlo subsets of the MCM I use the Kolomogorov-Smirnov statistic [Press et al., 1992]. This calculates the maximum distance D between the selected cumelative distribution function $S(x)$ (In our case the Logarithmic, Exponential and Power Law) of the data and the fitted model $P(x)$:

$$D = \max_{x \geq x_{min}} |S(x) - P(x)| \tag{4.3}$$

Using the MCM subsets from before it is seen that out of the three tested distributions, the MCM is best represented as a power-law distribution. Although this is not entirely within the chosen 5% significance, it is highly indicative that some aspects of the network are scale-free.

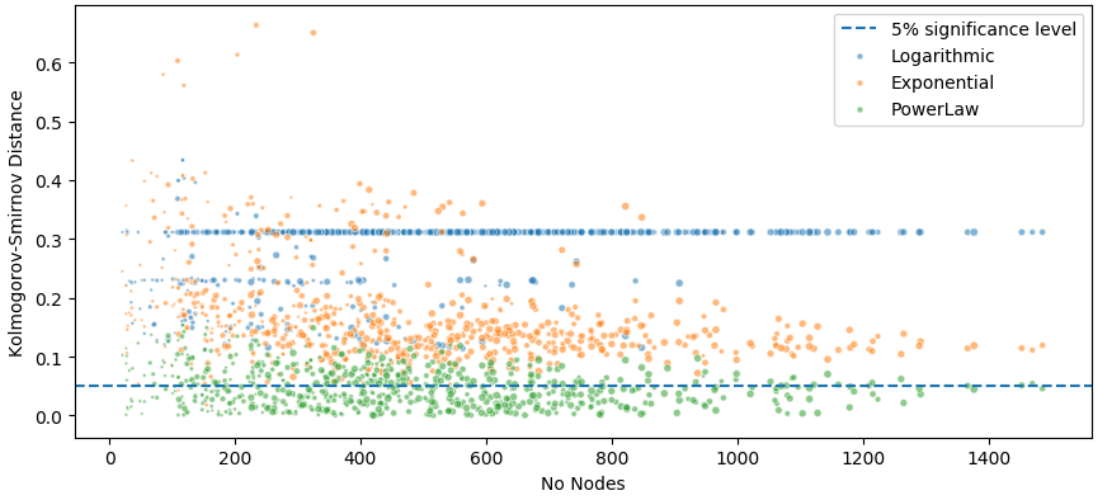


Figure 4.12: **Comparing the MCM subsets against a power law, logarithmic and exponential distribution.** The fit for different cumulative probability distributions of nodes in the MCM network is compared to determine the type of network hierarchy the chemistry follow. This is done by comparing the distance of the calculated distribution of data against a perfect one using the Kolomogorov-Smirnov test. The closer the two distributions are the better the fit.

4.4.4 Describing the MCM network

To conclude the MCM network exhibits both small world and scale-free (power-law) characteristics. This agrees with previous knowledge about the apparent network structure (branch and core - ref CH1/2). Here large primary emitted hydrocarbons produce branches of a hierarchical nature, as they are progressively broken down into smaller species. Since smaller species are then able to react with a much greater range of species, they then begin to form a tightly connected core, which exhibits many small-world features. This can be seen as the densely connected region within the graphs in CHAPTER !.

Having classified the MCM network type, the next section will look at how MCM based simulation results can be converted into the graph structure for more in-depth analysis, section 4.6.

4.5 Graph Construction methodology

Thus far we have only applied a qualitative analysis on the relationships between species in a mechanism. Although this can educate us about the chemistry within a specific system, often a quantitative value for the rate of reaction between different species is required when undergoing scientific evaluation or policy advice. To obtain such results a chemical mechanism is placed within an atmospheric model, initial concentrations are supplied and the chemistry is propagated forwards[10] in time. Currently, there exist three main model diagnostics which we may use to analyse the importance or role of a species from a simulation (model) output.

4.5.0.1 Concentration time series

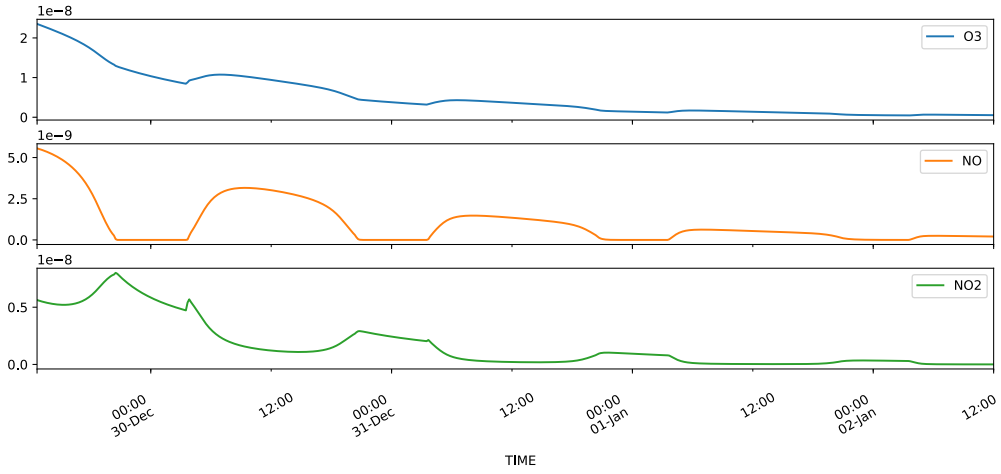
The simplest of these methods look at the abundance of a species at a specific point in the atmosphere - its concentration. As time moves forwards, chemicals within the atmosphere undergo a range of reactions which result in the making and breaking of bonds - thus the changing from one species to another.

Using the species concentration as a metric, we can map how it changes over time, and how in changing the initial concentrations of a simulation can produce different results. This can be useful for looking at a range of possible scenarios and evaluating the potential outcome after a pre-determined amount of time. An example would be through the use of policy-based simulations to predict changes in ...

Using a simple example from a Methane only subset of the MCM (Figure 4.13), it is possible to observe the inverse relationship between NO_2 and NO using only their concentration profiles. Here nitrogen

[10] Or backwards if the adjoint is used. (see section PAGERANK APPLICATIONS)

monoxide reacts with a RO_2 species to produce an RO and nitrogen dioxide. This then photolyses back to nitrogen oxide, releasing oxygen which may go on to form ozone (REF NOX CYCLE IN INTRO). The latter part of this reaction is dependant on photons and therefore can only occur during daytime (mostly).



$$NO \underset{h\nu}{\overset{RO_2}{\rightleftharpoons}} NO_2.$$

Figure 4.13: **A concentration time series from a simple methane-only simulation.** This is the simplest method for identifying changes in species within a model simulation. This multi-plot shows the changes in concentration profiles for all initialised species (NOx:10ppb; CH_4:20ppb; O_3:30ppb) following an initial 3 day spin-up to steady state.

4.5.0.2 Rate of Production and Loss

Analysing the concentration-time profiles allows the comparison of how 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 quantifies the specific reactions to which this may be attributed. Rate of Production Analysis (ROPA)[11] provides a method for establishing the total contribution from each reaction by calculating the change of concentration (concerning time) for the produced species - the instantaneous reaction Flux.

$$r_1 = \eta A + \omega B \xrightarrow{\kappa_1} C \qquad \text{Reaction 1} \tag{4.4}$$

$$f(C) = \frac{\delta C}{\delta t} = [A][B]\ \eta\omega \times \kappa_1 \qquad \text{Instantanious Flux } (\Gamma) \tag{4.5}$$

Here A, B and C are example species; [A],[B] and [C] are species concentrations; η and ω are rate coefficients and κ is the rate of the reaction.

[11] and loss

Using a sample simulation representative of the conditions within Beijing (an urban environment), we explore the reactions contributing to the production and loss of CH_3CO_3, Figure 4.14 at noon. The main reason for this specific example is that it can demonstrate how isolating a specific cause for the change within a species concentration may prove difficult in the context of atmospheric chemistry. Here we have many similarly weighted production and loss reaction, including that of peroxyacetyl nitrate (PAN) and nitrogen dioxide: $CH_3CO_3 + NO_2 \rightleftharpoons CH_3C(O)ONO_2$ (PAN). The reversible nature, coupled with its near-identical production and loss fluxes produce a very small net change within our species of interest (CH_3CO_3). Although this may be seen by calculating the cumulative flux between individual species, it is evident that simply looking at the concentrations or highest-ranking reaction fluxes may not be the best method of determining influence. To account for this we can look at how a change in one species can affect another using the Jacobian method.

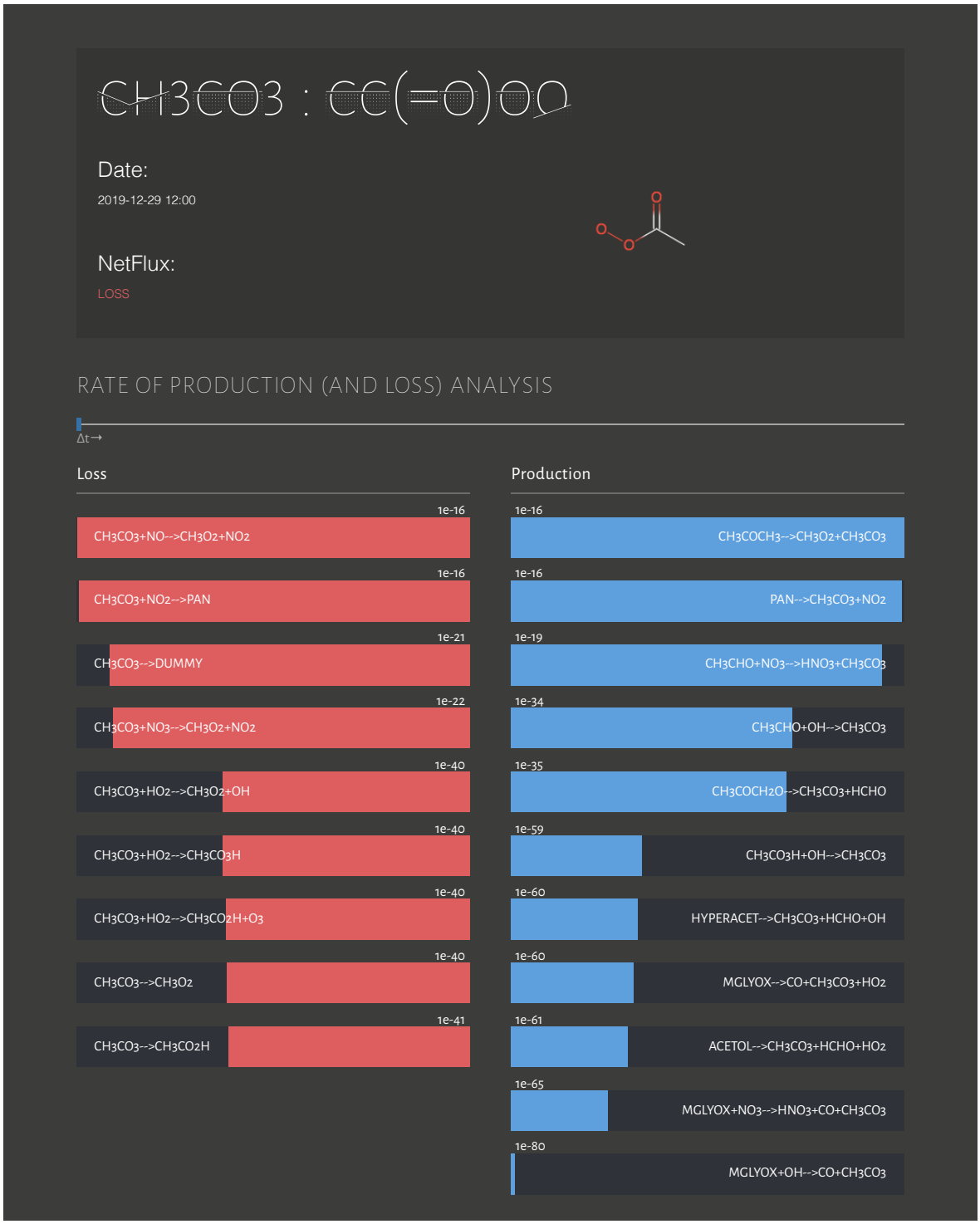


Figure 4.14: **Rate of production and loss analysis plot for CH_3CO_3 exhibiting a net loss (daytime).** An example ROPA plot from a simulation representing the chemistry within Beijing. This is used to identify the usefulness and weaknesses of using such a method.

4.5.0.3 The Jacobian

"The Jacobian [matrix] generalises the notion of gradient to describe the sensitivity to a vector" - Brasseur and Jacob [2017]. That this means is that in taking the partial derivatives of each reaction flux (e.g. from Equation 4.5), we can construct a representation of the influence each species has on itself - for example, the influence of species A on C and B on C (Equation 4.6-4.7).

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

$$\frac{\partial}{\partial B} \cdot \frac{\partial C_{r_1}}{\partial t} = \eta\omega A\kappa_1 \qquad \Gamma \text{ influence from B} \tag{4.7}$$

These partial equations can then be aggregated for all reactions that contain the two species - taking the effect of species B on species C, for example, produces Equation 4.8. Using these aggregate sums it is now possible to construct a pairwise relational matrix describing the influence each species has on every other species- Equation 4.9. This is known as the jacobian matrix and is what is used to propagate the chemistry within a simulation forwards in time.

$$\mathbf{J}_{C,B} = \frac{\partial f(C)}{\partial B} = \frac{\partial}{\partial B} \cdot \left(\frac{\partial \Sigma_{r_1}}{\partial t} + \frac{\partial \Sigma_{r_2}}{\partial t} + \cdots + \frac{\partial \Sigma_{r_n}}{\partial t} \right) \tag{4.8}$$

$$\mathbf{J}_{i,j} = \begin{bmatrix} \frac{\partial f_1}{\partial v_1} & \frac{\partial f_1}{\partial v_2} & \cdots & \frac{\partial f_1}{\partial v_n} \\ \frac{\partial f_2}{\partial v_1} & \frac{\partial f_2}{\partial v_2} & \cdots & \frac{\partial f_2}{\partial v_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial v_1} & \frac{\partial f_n}{\partial v_2} & \cdots & \frac{\partial f_n}{\partial v_n} \end{bmatrix}^{n,n}_{i,j=1} \tag{4.9}$$

4.5.1 Graph construction methodology for simulated data

Having covered the general definition of a Jacobian matrix and how it is constructed, we can now apply it to the context of mechanism analysis and comprehension. The first analogy that needs to be made is that for the flux, we have a the first differential of a specific reaction in time. If we consider the change in a species concentration as a 'displacement', we can think of the flux as its 'velocity'. Similarly, the Jacobian provides us with a description of how the individual flux of a species changes concerning the concentration (or displacement) or another species (the second-order partial differential). This is analogous to the acceleration of the object or particle we first displaced. In using the jacobian, we have constructed a relational matrix which outlines the effect a 1% change of a species has on all other species - a concept which is the foundation of the connectivity method (a mechanism reduction technique where all but essential and important species are removed), [Turányi and Tomlin, 2014].

Since the format of a jacobian is already in the form of a relational matrix, it can easily be converted to a weighted adjacency matrix, and then directly into the graph format. Since it only considers the aggregated influence between species, much of the work that would otherwise be needed to convert a mechanism into a graph format has already been done. To make use of the Jacobian matrix, several extraction algorithms were written for an updated version of the Dynamically Simple Model of Atmospheric Chemical Complexity (DSMACC) [DANDSMACC,DSMACC ref], as discussed in INTRODUCTION. Here we edit the kinetic pre-processor output, [**?**] to release the values of the Jacobian Matrix and return them at each model timestep for analysis. The process for how this is done is described in subsection 4.5.2.

A note on using the Flux instead of the Jacobian

Depending on the model setup or the users' capabilities, extraction of the jacobian matrix for each timestep may not be possible. In many cases, the reaction rates and concentration may still be available, allowing for the calculation of reaction fluxes throughout the simulation. If this is the case the total flux can be calculated using the method described in Equation 4.5. From this, an edge-weighted by a reaction flux can be created from every reactant to each product. This generates a multi-graph[12] which may be simplified by taking the net flux value for all edges between two nodes.

However, the potential for human/coding error, additional simplification and a non-explicit definition of the contribution of each species make the use of a Jacobian much more efficient in network generation from a chemical mechanism.

[12] A graph with multiple edges between nodes

4.5.2 A practical Example using the MCM

Taking a single equation from the MCM we may calculate the jacobian relationships between species and convert them into a graph. A randomly chosen ethane reaction (Equation 4.10) from a simple mechanism was chosen. It must be noted that in general it is unusual in the MCM that alkyl radicals react rapidly and extremely well with O_2 to from stabilised peroxy radicals, [Jenkin et al., 1997]. In general, the reaction would consist of the following two steps: $C_2H_6 + OH \xrightarrow{\kappa_1} C_2H_5\cdot + H_2O$ and $C_2H_5\cdot + O_2 \longrightarrow [\kappa_2]\, CH_2H_5O_2$.

$$C_2H_6 + OH \xrightarrow{\kappa_3} C_2H_5O_2 \tag{4.10}$$

For simplicity in this example, this will be the only equation for our mechanism. The resultant Flux Equation 4.11 and resultant Jacobian Equation 4.12 may be calculated.

$$\Gamma = [C_2H_6][OH]\kappa_1 \tag{4.11}$$

$$\mathbf{J}_{i,j} = \begin{bmatrix} \frac{\partial f_{[C_2H_6]}}{\partial t\ \partial[C_2H_6]} & \frac{\partial f_{[C_2H_6]}}{\partial[OH]} & \frac{\partial f_{[C_2H_6]}}{t\partial[C_2H_5O_2]} \\ \frac{\partial f_{[OH]}}{\partial t\ \partial[C_2H_6]} & \frac{\partial f_{[OH]}}{\partial t\ \partial[OH]} & \frac{\partial f_{[OH]}}{\partial t\ \partial[C_2H_5O_2]} \\ \frac{\partial f_{[C_2H_5O_2]}}{\partial t\ \partial[C_2H_6]} & \frac{\partial f_{[C_2H_5O_2]}}{\partial t\ \partial[OH]} & \frac{\partial f_{[C_2H_5O_2]}}{\partial t\ \partial[C_2H_5O_2]} \end{bmatrix}_{i,j=1}^{3,3} \tag{4.12}$$

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 4.13.

$$\mathbf{J}_{i,j} = \begin{bmatrix} \frac{\partial f_{[C_2H_6]}}{\partial t\ \partial [C_2H_6]} & -2 \times 10^{-7} & 2 \times 10^{-7} \\ -0.1 & \frac{\partial f_{[OH]}}{\partial t\ \partial [OH]} & 0.1 \\ & & \frac{\partial f_{[C_2H_5O_2]}}{\partial t\ \partial [C_2H_5O_2]} \end{bmatrix}_{i,j=1}^{3,3} \tag{4.13}$$

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 4.15.

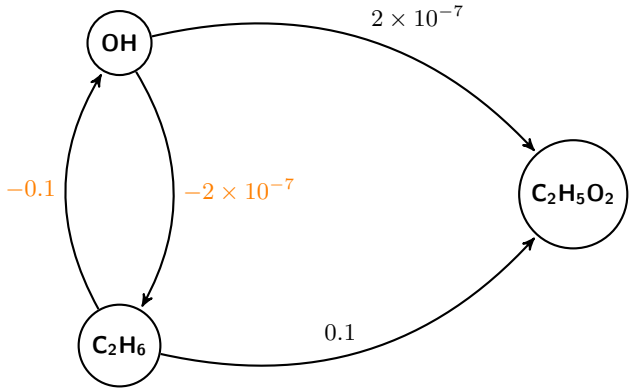


Figure 4.15: A graphical representation of Equation 4.13 derrived from the Equation 4.10

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 4.13 inherently holds such property and can be directly translated to produce a graph, Figure 4.15. 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 4.16.

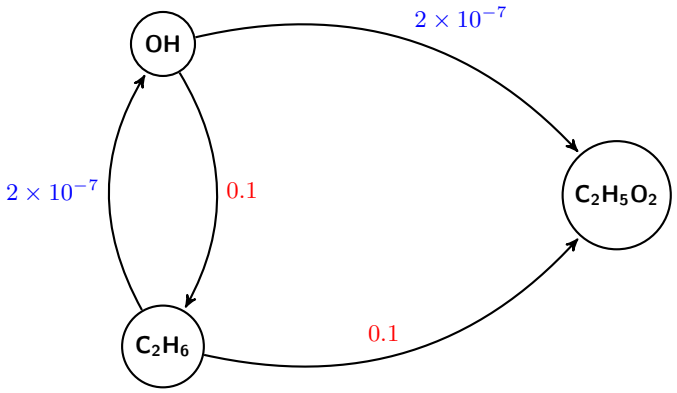


Figure 4.16: Reversing the directions on negatively weighted edges from Figure 4.15

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 4.17. 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.

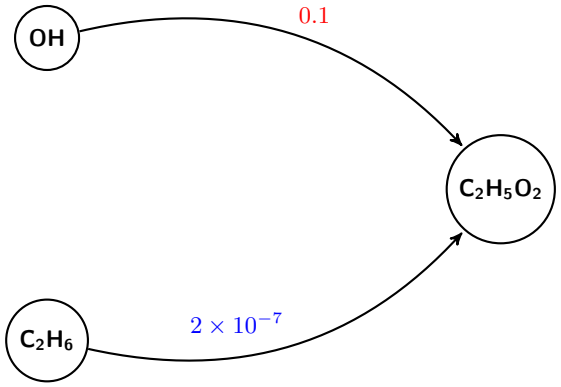


Figure 4.17: Simplifying Figure 4.16

4.6 Case study Example

In this section the centrality metrics discussed in section 4.3 are applied to a range of senarios. These range from polluted urban environments such as London [REF] and Beijing [REF], to marine and terrestial forrest- Cape Verde REF and Borneo REF. We determine the main drivers for the chemistry and compare the species which are important across each simulation.

4.6.1 Establishing Initial Conditions from observational data

Within experimental data assimulation it is not uncommon to face problems which resut in unreliable or missing data. These can range from anything as little as measuring below the instrument sensitivity

to powercuts and equiptment damage/theft from the local wildlife. This can result in problems when analysing the results and combining them to create a simulation of the chemistry for that environment.

To overcome this, traditionally a combination of data filtration, smoothing and interpolation is required. Although it is possible to fit a diurnal profile, through itereative methods of comparison, and cubic splines, a much simpler way would be to use an Multi Layer Perceptron Regressor model (MLPRegressor) as provided by sklearn, [**?**]. This is described below.

4.6.1.1 The origin of Artificial Neural Networks

The concept of a neural network originated witin the field of neuroscience. In in bilogical neurons, signals are sent through the use of electrical impulses using their synapses. When a sufficient number of signals are recieved within a short timeframe, a neurone will respond, often firing a range of its own signals. Using this as a foundation,McCulloch and Pitts [1943] presented a computational model of the biological neuron - the artifical neuron. This has a series of binary inputs, and produces a single binary output. This idea was later imporoved with the invention of the perceptron - a linear classifier which classifies categories by separating them with a straight line. Invented by Rosenblatt [1958], this was popularised as a device representative of a modern day shallow neural network - [John Hay, 1960], Figure 4.18. Unlike the artificial neuron however, the perceptron is able to take non-binary (numerical) inputs of an associated weight which allows for the computation of simple linear binary classification. Much like Logistic regression, the preceptron produces a positive or negative classification based on a certain threshold[13].

[13]It is worth noting that while a Logistic Regression classifier can output a class probability, the use of a hard threshold means that this is not done within the preceptron algorithm [Géron, 2017]

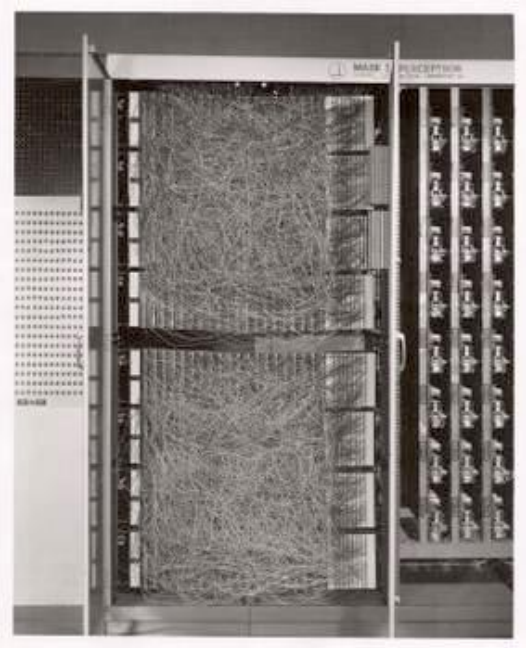


Figure 4.18: **The Mark 1 perceptron** Both software and hardware are different manifestations of a flow chart. The perceptron hardware accomplished what is now done using software. Source: Cornell [2020]

4.6.1.2 The Multi Layer Perceptron

Limitations of ther perceptron include the classifiction of complex patterns such as the XOR problem (where a category appears between two other categories e.g. 1|0|1 - this cannot be classified by a single linear split). In taking inspiration from nature, Figure 4.19, it is possible to overcome this with the use of multiple layers. This creates an a deep ($>$ 2 two hidden (non-input) layers of perceptrons[14]) artificial neural network (ANN)

The multi layer perceptron (MLP) model now represents a simple feed-forwards network, much like a decision tree. However unlike a decision tree, the MLP ANN is able to describe the probability a branch is taken using non-linear activation (threshold) functions. These are discussed in detail as part of SECTION NEXT CHAPTER. The weighting thresholds for each neuron are then calculated by backwards propagation of results through the network until a suitably good result is produced.

> ***Example analogy:*** *Back propagation can be likened to the iterative calibration of scientific instrumentation. In the field of atmopsheric chemistry laser induced flourecence is used to calculate species concentrations and reaction rates within the troposphere, [Dillon et al., 2006; Bloss et al., 2004]. Here the frequency of a laser can be adjusted in contrast with a known target (e.g. an amount of SO_2) to produce a responce curve showing where the maximum resonance occurs.*

[14] These are sometimes refered as Linear Threshold Units.

Similarly a neural network can be 'trained' (calibrated). This is done through the use of a 'training dataset' - a set of input-output pairings which represent a random selectaion of 2/3rds of the total dataset. Next the neurons within each layer (similar to the potentiometer dials on an instrument) are adjusted in sequence through the layers to match the known result (a standard of known concentration) to the input values provided. This process is repeated until for a number of iterations, or until a sufficently 'good' prediction is attained for the entire training dataset (early termination). The power of ANNs comes from the ability to adjust neuron thresholds whilst moving both forwards and backwards through the network (Note: predictions of a MLP are still only passed forwards). Finally model perfomance is evaluated against the remaining 1/3rd of the total dataset.

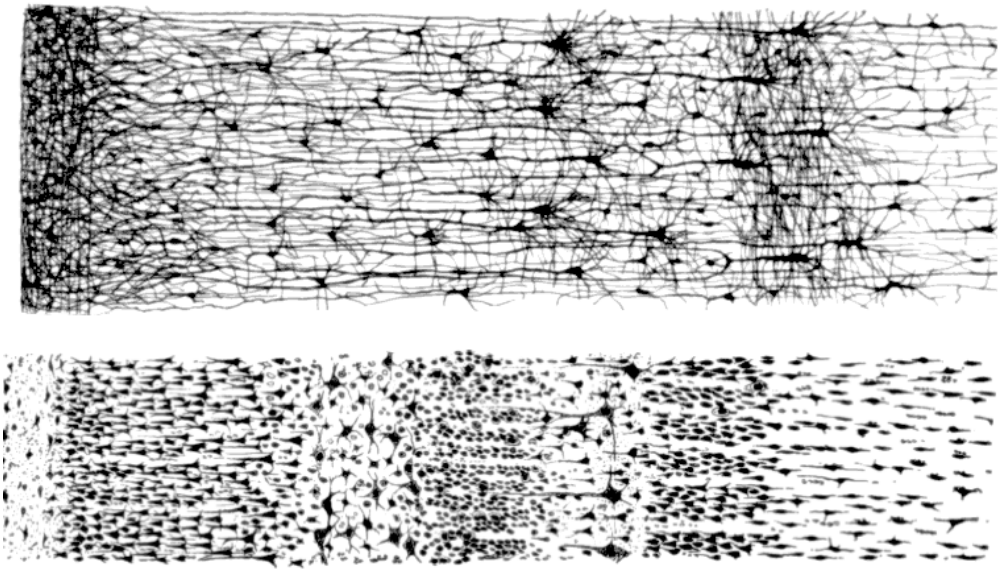


Figure 4.19: **The Human Cortex - A biological neural network.**. A vertical cross section of the human cortex between an adult (top) and 1.5 month old infant (bottom) showing a layer like structure with a change in depth (left to right). Source: Cajal [2020]

4.6.1.3 Application of the MLPRegressor to observational data

In the application of any type of machine aided algorithms it is important to evaluate the results provided. In this section the results of 12 years of data collected as part of the [CAPE VERDE CAMPAIGN] are shown (these contain measurements spanning the entirety of 12 years, which produce the clearest tests for the algorithm). A MLPRegressor of 10 hidden layers, and a hyperbolic tan (tanh) activation function is used. Additionally the limited-memory Broyden–Fletcher–Goldfarb–Shanno (l-BFGS) solver (a quasi-newton method which minimises the inverse of the Hessian matrix[15] to steer

[15]The hessian is square matrix of second-order partial derivatives of a scalar-valued function/field describing the local curvature of a function (of many variables).

through space and obtain a solution) and an adaptive learning rate[16] is used.

Input of the regressor is in the from of a month and a hour, to represent each measurement. This allows it to find not only daily trends, but also seasonal trends within the data. Once trained the regressor is then used to predict a diurnal profile for each month based on the observational data provided. For simplicity $\log_{10}$ values of the concentrations obtained have been used. To validate the results, the predicted MLPRegressor line is compared to a transparent scatterplot for all the results. In addition to this a boxplot showing the IQR, median and mean (green line) plotted alongside to evaluate the predictor output.

In providing the MLPRegressor with both month and hour inputs, the data is not only fitted hourly (a diurnal avarage), but also across the seasonal/monthly cycles. This accounts for the variation between years and datasets. Since $\log_{10}$ values of the concentrations are used, species such as ozone (Figure 4.20) which for the Cape Verde dataset (clean air) do not change more than one order of magnitude, the effects of neigbouring months, which shift the diurnal away from the mean (the green line on the boxplot), can be seen. However since this is overall a small change, and the diurnals lie within the inter quartile range, they still provide an adequate approximation. NO (Figure 4.21) on the other hand has a concentration change of several orders of magnitude. Here a distinct daytime peak is seen and is centred around a sesonally consistent mean value of the data. Here the multi-magnitude change in concentration also provides an effective silhouette of the data to which we may compare the fitted line. Finally the plots of NO_2 and iso-Pentane (Figure 4.22-4.23) vary both in diurnal magnitude and seasonally. Within these plots, changes in the data in the january and december months produce deceptively misleading results. Here although the diurnals are not symmetrix, they fit well within the median,mean and interquartile range values, as well as the general data silhouette behind them. This siggests that it is a property of the data that we are fitting, and not that the regressor is producing incorrect results. It is however noted that for a more accurate seasonal prediction, periodic boundary conditions should be employed in the training dataset, where an additional two months are added before January and after December.

Since I shall be using only a noon estimate from the summer region, this does not affect any of the results.

[16]Each time the model improvement fails to decrease the learning loss, the learning rate is reduced by 1/5. This means smaller jumps are made towards the curve peak.

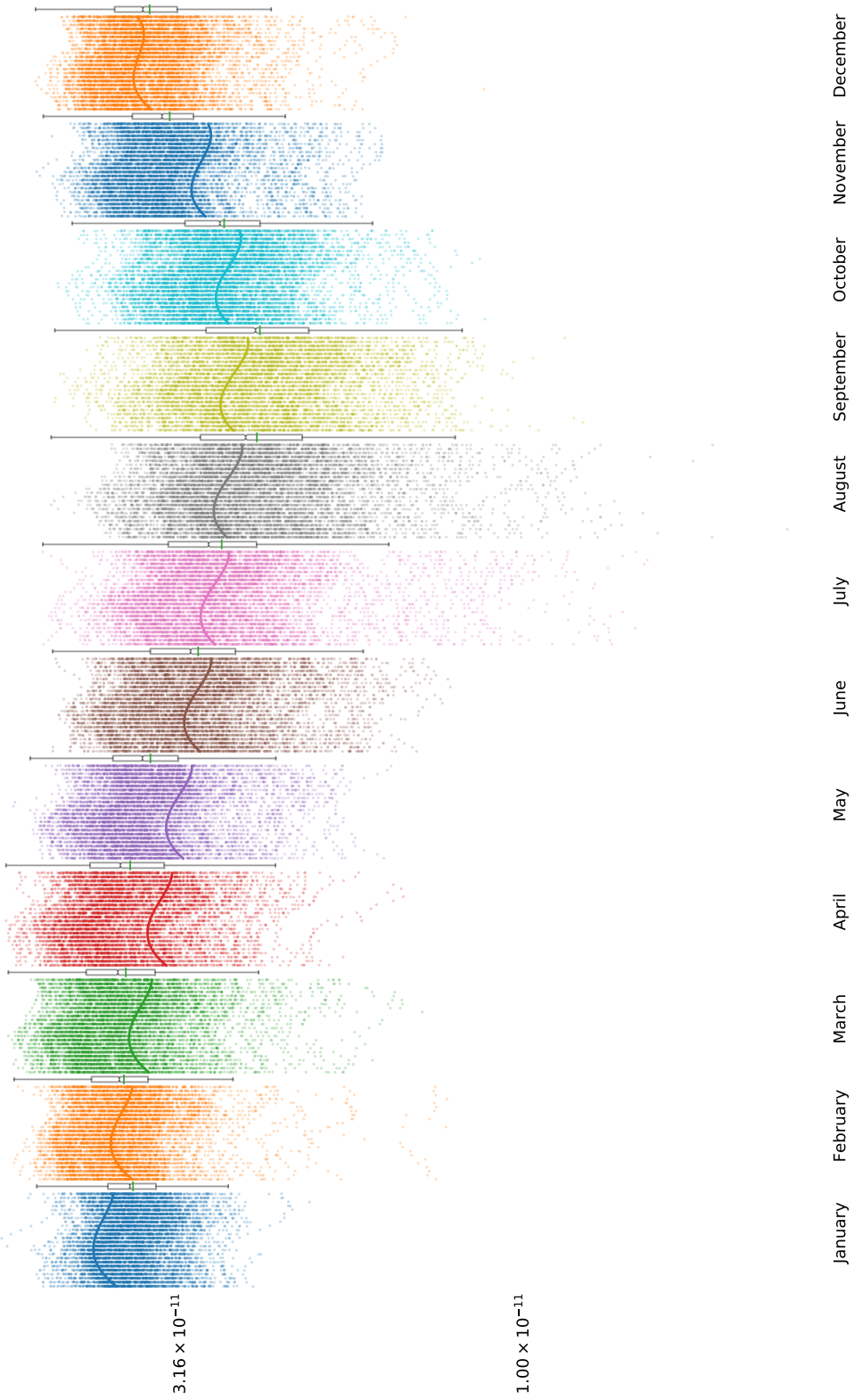


Figure 4.20: **Cape Verde MLP predicted and observational data of Ozone.** Each segment represents data from a different month. Within each month segment exists 24 hour segments to create a diurnal. Observational concentrations are plotted in the form of a translucent scatterplot and summarised using the boxplot on the right of the month segment. MLP predicted results are shown using the solid lines. Concentration in mixing ratio.

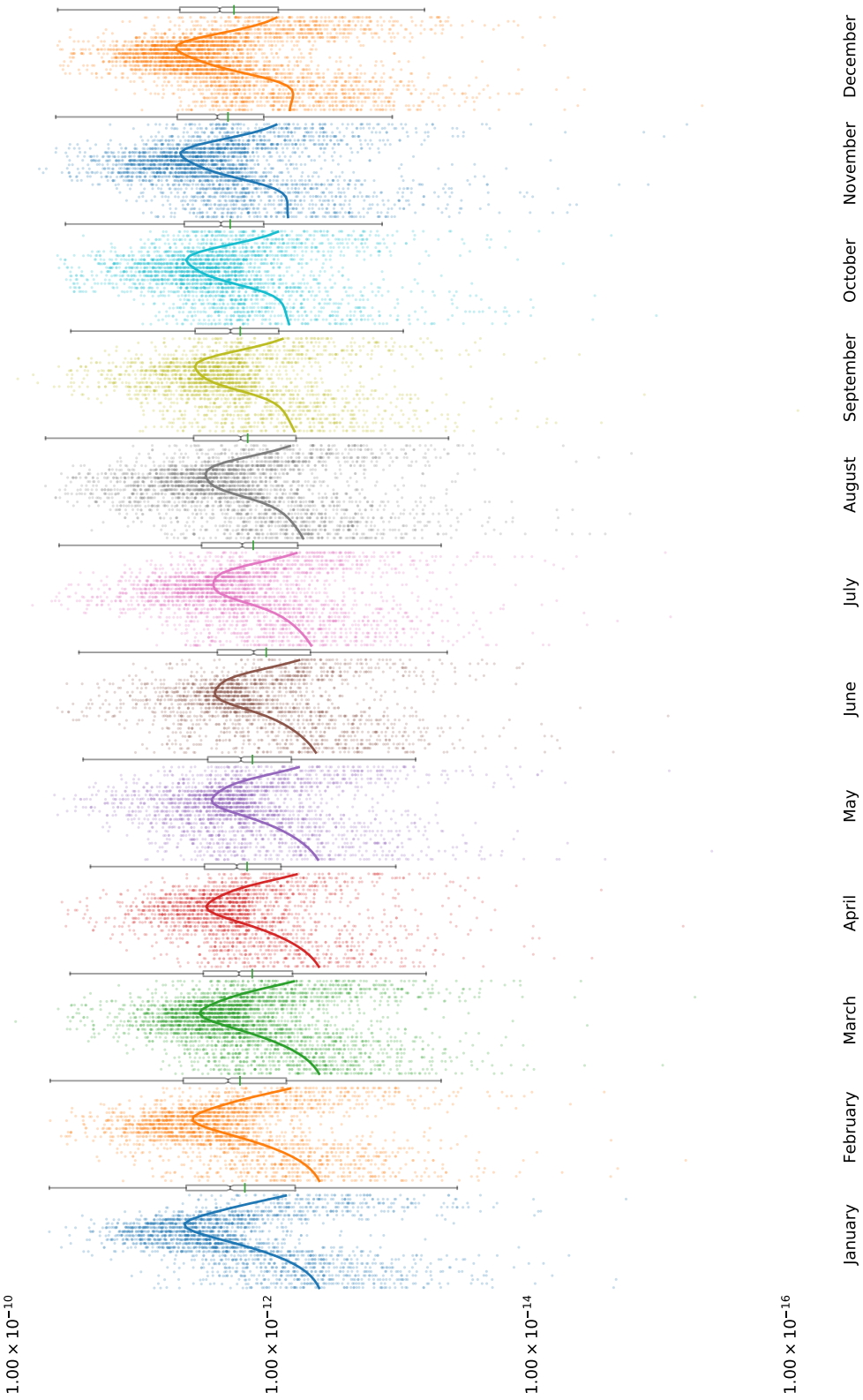


Figure 4.21: **Cape Verde MLP predicted and observational data of NO.** Each segment represents data from a different month. Within each month segment exists 24 hour segments to create a diurnal. Observational concentrations are plotted in the form of a translucent scatterplot and summarised using the boxplot on the right of the month segment. MLP predicted results are shown using the solid lines. Concentration in mixing ratio.

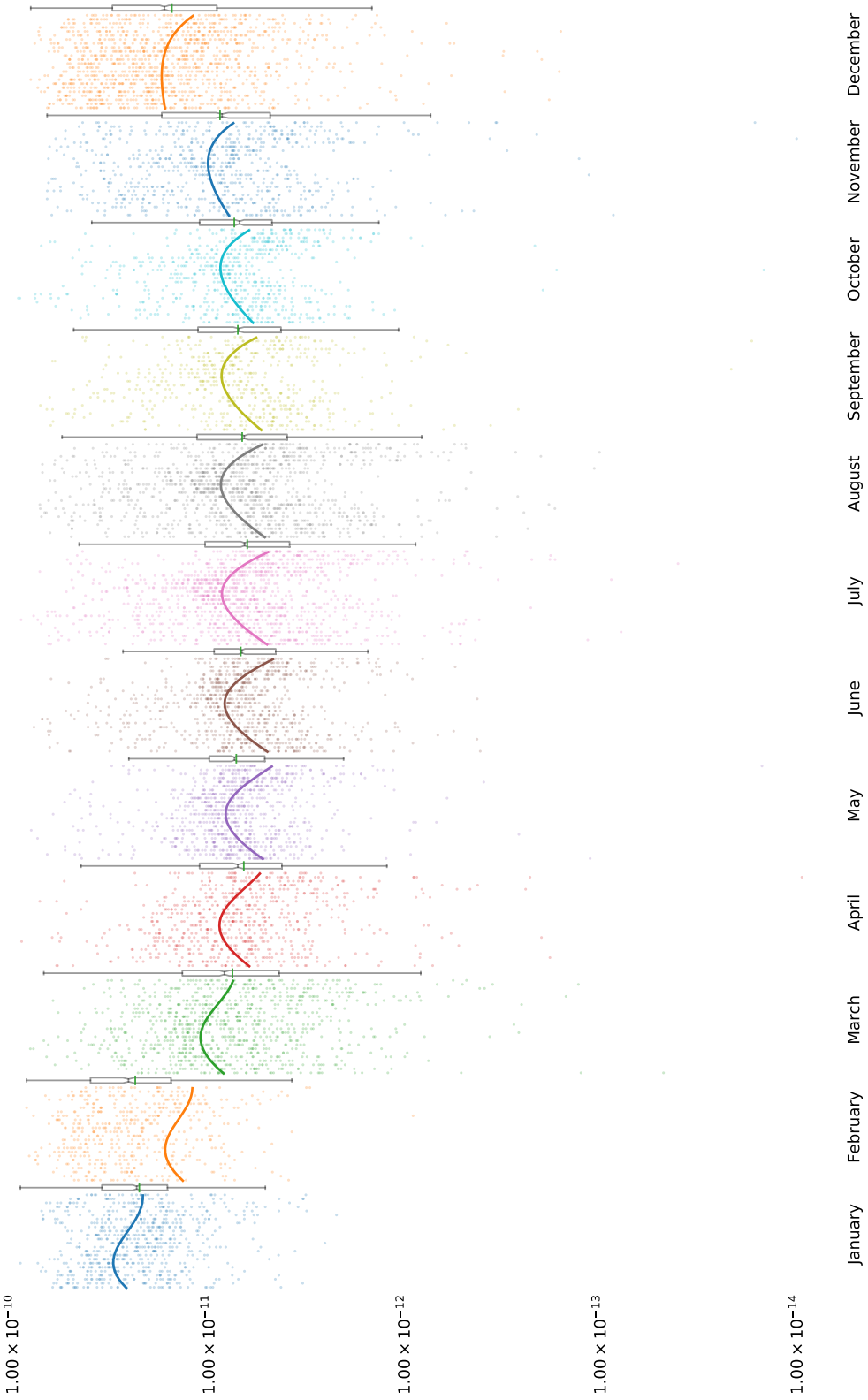


Figure 4.22: **Cape Verde MLP predicted and observational data of NO_2.** Each segment represents data from a different month. Within each month segment exists 24 hour segments to create a diurnal. Observational concentrations are plotted in the form of a translucent scatterplot and summarised using the boxplot on the right of the month segment. MLP predicted results are shown using the solid lines. Concentration in mixing ratio.

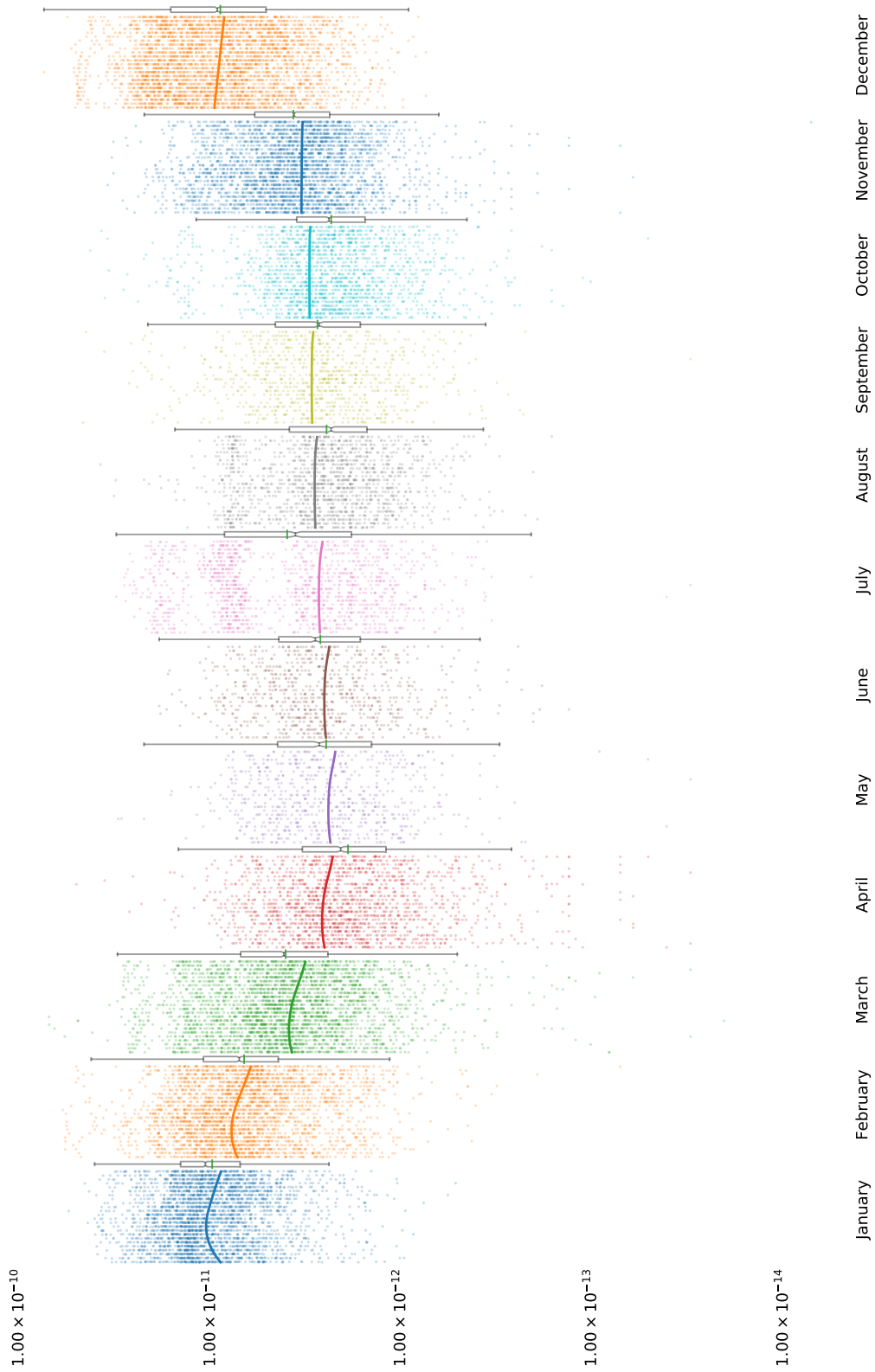


Figure 4.23: **Cape Verde MLP predicted and observational data of iso-Pentane.** Each segment represents data from a different month. Within each month segment exists 24 hour segments to create a diurnal. Observational concentrations are plotted in the form of a translucent scatterplot and summarised using the boxplot on the right of the month segment. MLP predicted results are shown using the solid lines. Concentration in mixing ratio.

4.6.1.4 Generation of the ICS file and simulation run

To generate the initial conditions for a model run, the MLP regressor is used to predict a representative species concentration at noon. As campaigns run at different times in the year, an error of $\pm 2 months$ around June exists.

The predicted initial conditions in Table **??** are used to initiate a box model simulation. To spin these up the initial concentrations are reset at noon each day. When the simulation reaches a fractional difference of less than 0.001, the model is left to run unconstrained for 5 days. Observation of the concentration profiles of each run (Figure **??**-**??**) show us how RO_2, HO_x, NO_x and Ozone change over time. Within these it can be seen that RO_2 takes longer to stabilise in the polluted environments (London and Beijing) than those with cleaner air (Cape Verde and Borneo). Using the concentration profiles as a guide the instantaneous values for the simulation results are taken at noon, 3 days after the end of the spinup period.

4.6.1.5 Extracting the required results

Model diagnostics such as concentration and the net flux passing through a species may be extracted directly from the DSMACC box model. These provide the baseline comparison against which we shall compare the results of graph metrics. The concentration will tell us which species are most abundant, and the net-flux tells us which are the fastest changing (either as a production or loss).

In addition to this the jacobian matrix at the final timestep shall also be extracted. This is then manipulated to remove negative links (AS IN SECTION XX) and normalised. Using the normalised jacobian as an adjacency matrix, a graph on which the centrality metrics can be applied is generated. The result of these is compared in the next section.

4.6.1.6 Unifying the results

Since each metric compared can produce results on a different scale, there needs to be a method of normalising the results to achive parity between the range in which species are ranked. To do this all values within a group can be scaled between zero and unity, where 1 is the highest ranked and 0 is the lowest. For entries such as concentration or flux, which span several orders of magnitude, the $\log_{10}$ of the values shall be taken before they are scaled.

Species	Beijing(APHH)	Borneo(OP3)	London(ClearFlo)	CapeVerde
LAT	39.9	0.96	51.0	16.5
LON	116.3	114.5	0.00	23.4
O_3	6.883e-08	8.939e-09	3.819e-08	2.629e-11
NO	1.660e-09	2.668e-14	2.350e-09	2.358e-12
NO_2	1.226e-08	1.081e-13	7.445e-09	8.447e-12
HCHO	4.472e-09		1.119e-08	
C_2H_6	3.163e-09	7.315e-10	2.133e-09	4.539e-10
C_2H_4	1.004e-09	1.152e-10	4.893e-10	2.481e-11
C_3H_8	3.019e-09	1.924e-10	1.128e-09	1.728e-11
C_3H_6	1.335e-10	1.333e-11	1.784e-10	9.343e-12
IC_4H_{10}	6.412e-10	8.742e-11	5.142e-10	2.486e-12
NC_4H_{10}	1.593e-09	5.698e-11	1.058e-09	4.481e-12
C_2H_2	1.058e-09	1.825e-10	3.018e-10	1.848e-11
TBUT2ENE	4.198e-11		1.815e-11	
CBUT2ENE	4.454e-11		1.305e-11	
IC_5H_{12}	1.047e-09	2.883e-11	7.424e-10	3.470e-12
NC_5H_{12}	4.650e-10	2.090e-11	2.792e-10	2.513e-12
TPENT2ENE	3.939e-11			
CPENT2ENE	3.982e-11			
NC_6H_{14}	2.057e-10	6.437e-12	6.357e-11	
C_5H_8	7.134e-10	1.957e-09	1.640e-10	
NC_7H_{16}	7.905e-11		5.222e-11	
BENZENE	4.045e-10		1.137e-10	7.682e-12
NC_8H_{18}	3.091e-11		1.442e-11	
TOLUENE	6.767e-10		3.205e-10	3.121e-12
EBENZ	3.115e-10		6.017e-11	
OXYL	1.677e-10		5.049e-11	
CH_3CHO	4.783e-10		4.095e-09	
C_2H_5OH	4.655e-09		3.125e-09	
CH_3COCH_3	3.328e-09		2.924e-09	
NC_9H_{20}	1.336e-11		7.922e-11	
$NC_{10}H_{22}$	1.062e-12		1.602e-10	
α-PINENE[17]	7.341e-11	15e-11	1.105e-10	
LIMONENE	5.836e-11	1.351e-10	3.566e-11	
$PXYL^+MXYL$[18]	4.943e-10			
IPBENZ	4.567e-10			
PBENZ	3.996e-10			
HONO	6.479e-10		4.109e-10	
MACR		6.948e-11	1.862e-11	
$PENT_1ENE$			2.383e-11	
MVK			2.091e-11	
NPROPOL			2.883e-10	
NBUTOL			4.535e-10	
STYRENE			2.241e-11	
MEK			5.494e-11	
C_3H_7CHO			9.534e-12	
C_4H_9CHO			1.865e-11	
$C_5H_{11}CHO$			1.201e-11	
CYHEXONE			9.790e-12	
BENZAL			1.510e-11	
PAN			1.791e-10	

Table 4.4: The initial conditions created from the MLPRegressor prediction of observational data.

[18]This is written as ?-pinene in the merged CEDA dataset for the Borneo OP3 campaign. This is due to character conversion errors.

[18]The concentration for these is split evenly between both species

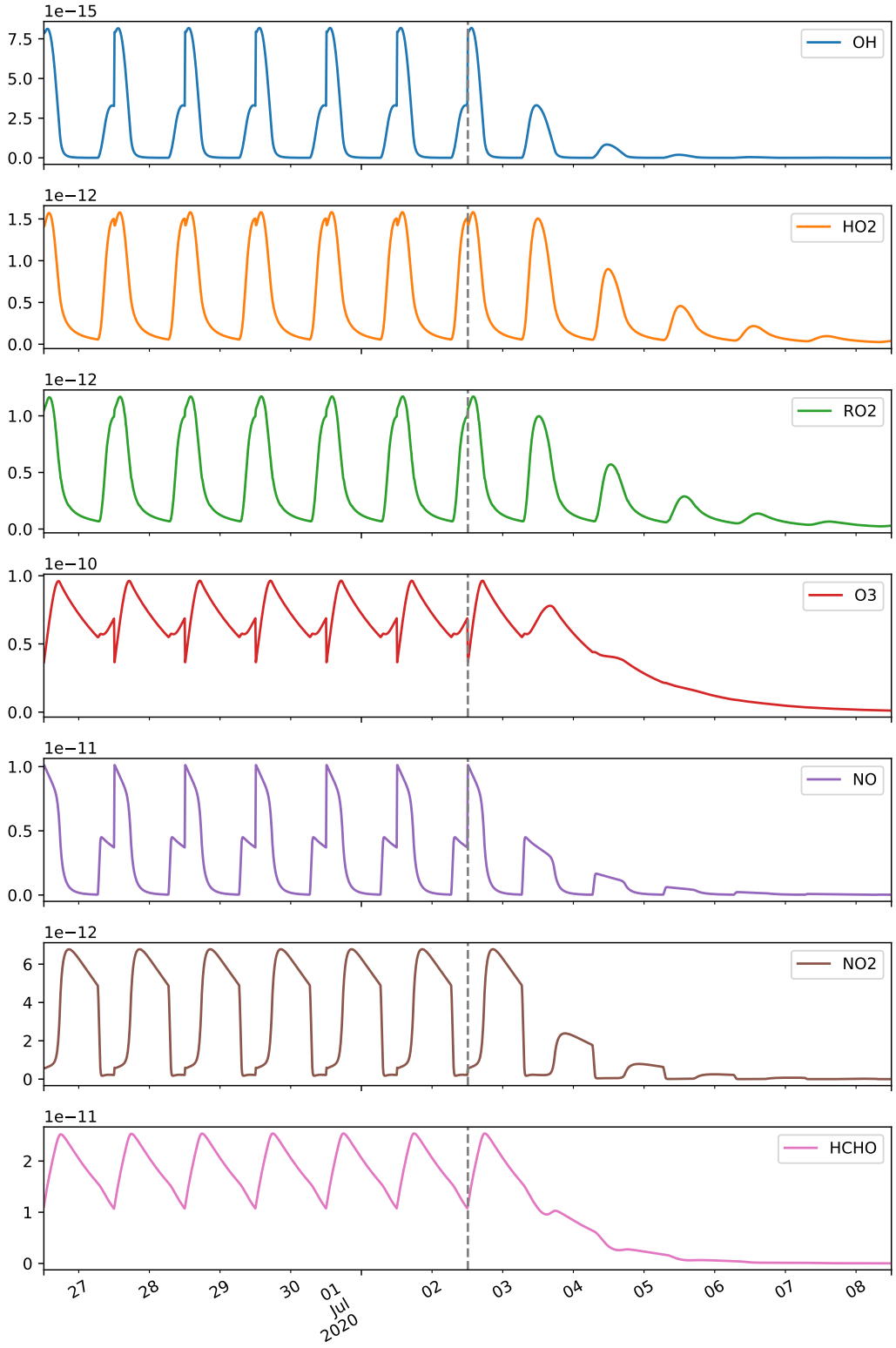


Figure 4.24: **The concentration profile for CapeVerde.**This shows a the change in concentration over time for HO_x,NO_x,Ozone and RO_2 species for a simulation run generated by the mlpregressor. Left of the dashed line shows the last 6 days of spinup, where the intial concentrations are reset at noon each day until the species fractional difference is less than 0.001 .

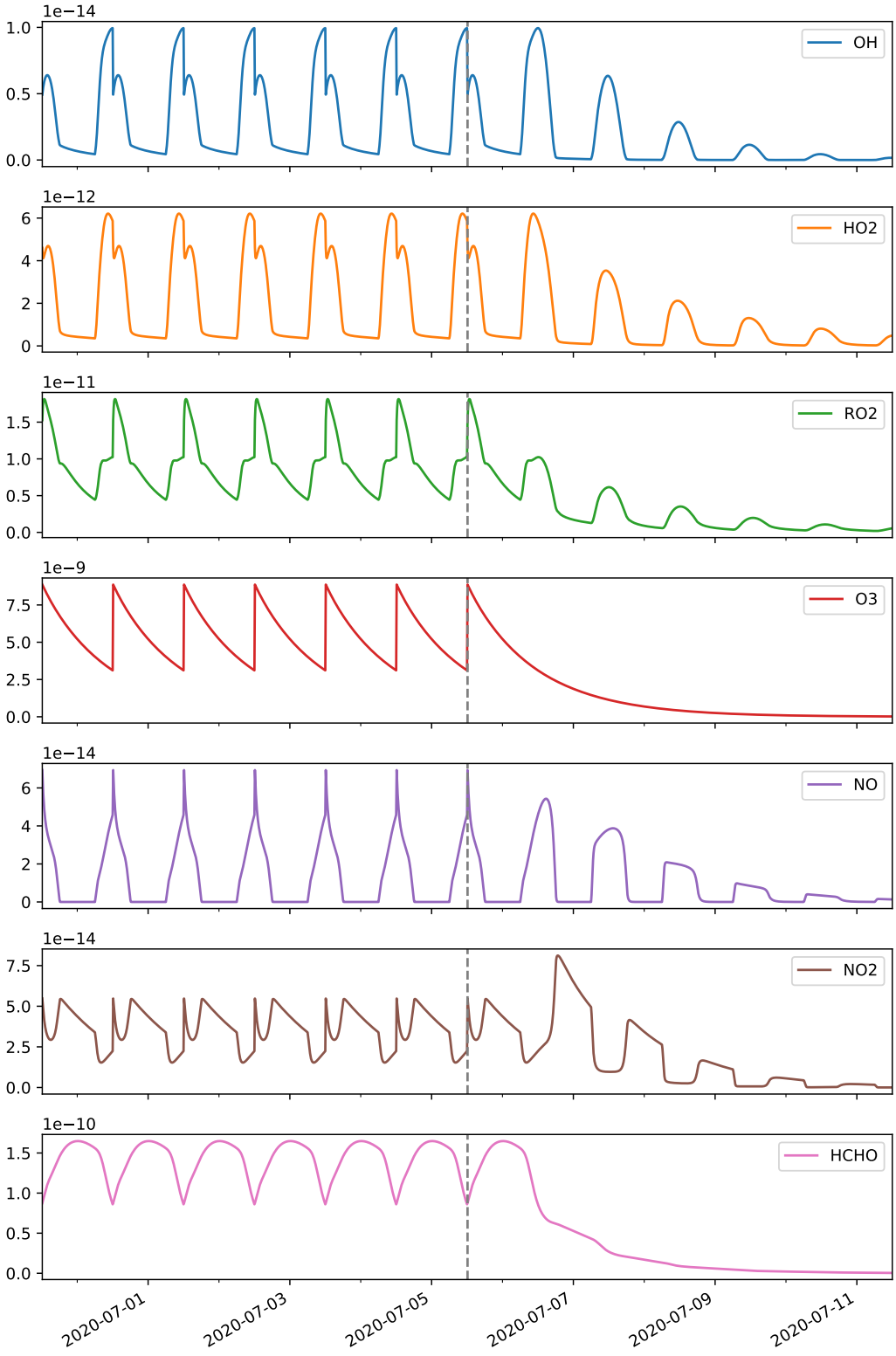


Figure 4.25: **The concentration profile for Borneo.**This shows a the change in concentration over time for HO_x,NO_x,Ozone and RO_2 species for a simulation run generated by the mlpregressor. Left of the dashed line shows the last 6 days of spinup, where the intial concentrations are reset at noon each day until the species fractional difference is less than 0.001 .

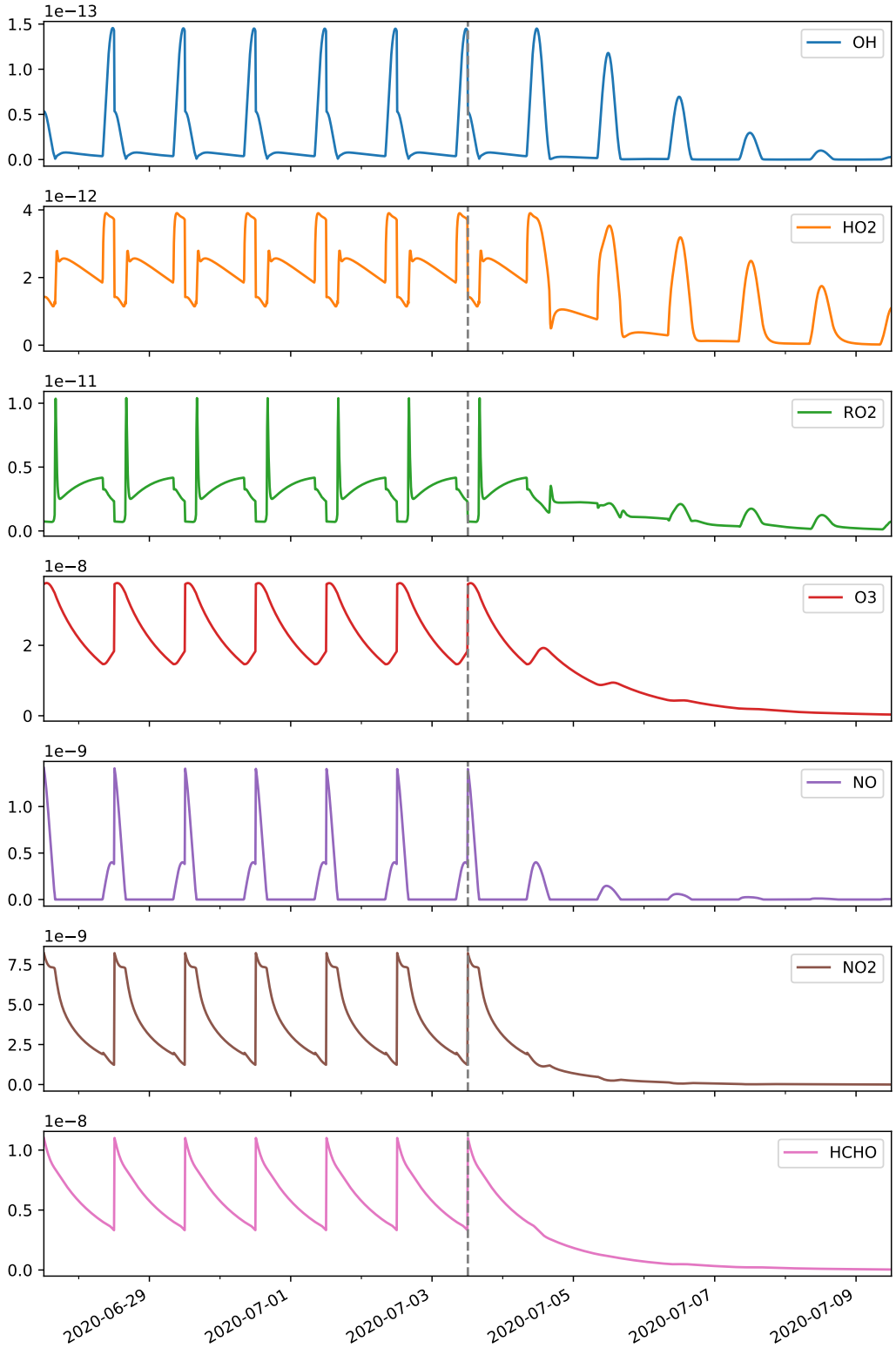


Figure 4.26: **The concentration profile for London.**This shows a the change in concentration over time for HO_x,NO_x,Ozone and RO_2 species for a simulation run generated by the mlpregressor. Left of the dashed line shows the last 6 days of spinup, where the intial concentrations are reset at noon each day until the species fractional difference is less than 0.001 .

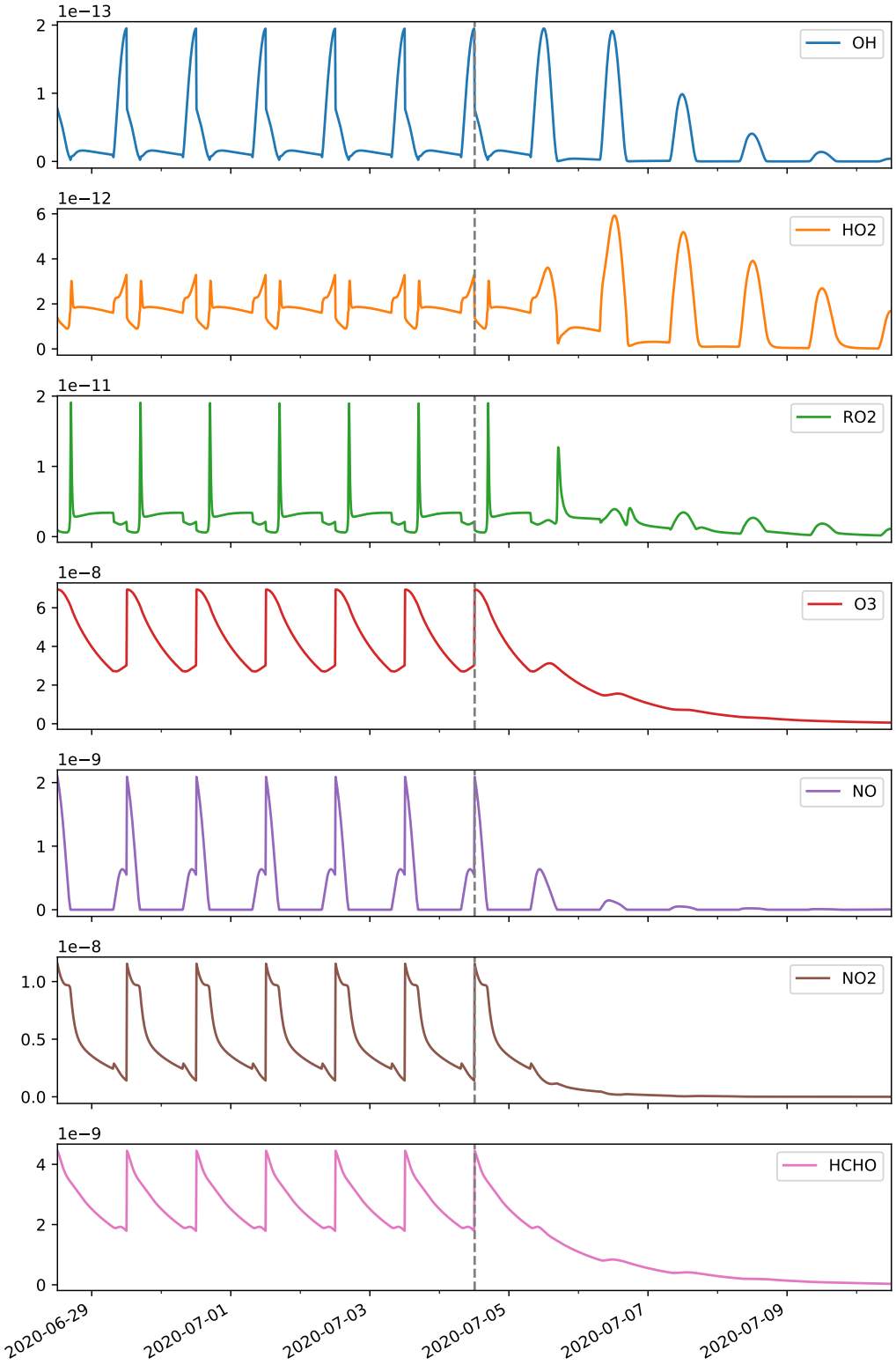


Figure 4.27: **The concentration profile for Beijing.**This shows a the change in concentration over time for HO_x,NO_x,Ozone and RO_2 species for a simulation run generated by the mlpregressor. Left of the dashed line shows the last 6 days of spinup, where the intial concentrations are reset at noon each day until the species fractional difference is less than 0.001 .

4.7 Comparing Results

This subsection presenets the comparison between traditional model diagnostic methods and the different graph metrics. It shall also provide an inter-senario comparison between the different types of chemistry which exists in each simulation.

Since each simulation consists of several thousand species, it is infeisable to attempt to directly compare them for each centrality metrics. This is especially true for species that may not be consistently important across all metrics. To overcome this a computational method for extracting keywords from a corpus of documents, Term Frequency - Inverse Document Frequency (TF-IDF), will be used.

4.7.0.1 The inner workings of TF-IDF

TF-IDF is a numerical statistuc used in text natural language processing and text mining to identify the importance of a word with regard to its context. It provides a value for the frequency a word appears within a section, offset by the number of times it appears in other sections - It is for this reason that 83% of text recommender systems in digital libraries use TF-IDF, [Beel et al., 2016].

In [Ellis, 2019] I applied this to the chapters of frankenstein, and found the keywords extracted almost exactly replicated those from the synoptic description of the novel. Although TF-IDF is a text mining procedure, the algorithm itself is mathematical, meaning that it may be applied to our diagnostic dataset. The working of the algorithm are discussed below.

Term Frequency

The TF from the algorithm name stands for term frequency. This is an analysis of the number of times a word exists within a dataset. There are several ways in which this can be done, these are:

- **Raw Count** - The *number of times* a word exists within the document.
- **Boolean/Logistic** - $True$ if the word exists, false otherwise.
- **Adjusted for Document Length** - $word\ frequency/total\ number\ of\ words$
- **Log Scaled** - $\log(1 + frequency)$

As the scaled values for each item are taken, we can liken our results to the 'Adjusted for Document length' equation and use the scaled ranking value for each group respectively.

Inverse Document Frequency

Inverse document frequency tell us how much information a word provides with respect to a certain context. Whilst a word may be used extensively throughout the corpus (term frequency) we are often interested in words which ocuur often only in a single section. This is one of the reason TF-IDF is useful in the extraction of keywords from a document.

To complete the TF-IDF equation, the term frequency and inverse document frequency terms are multiplied together.

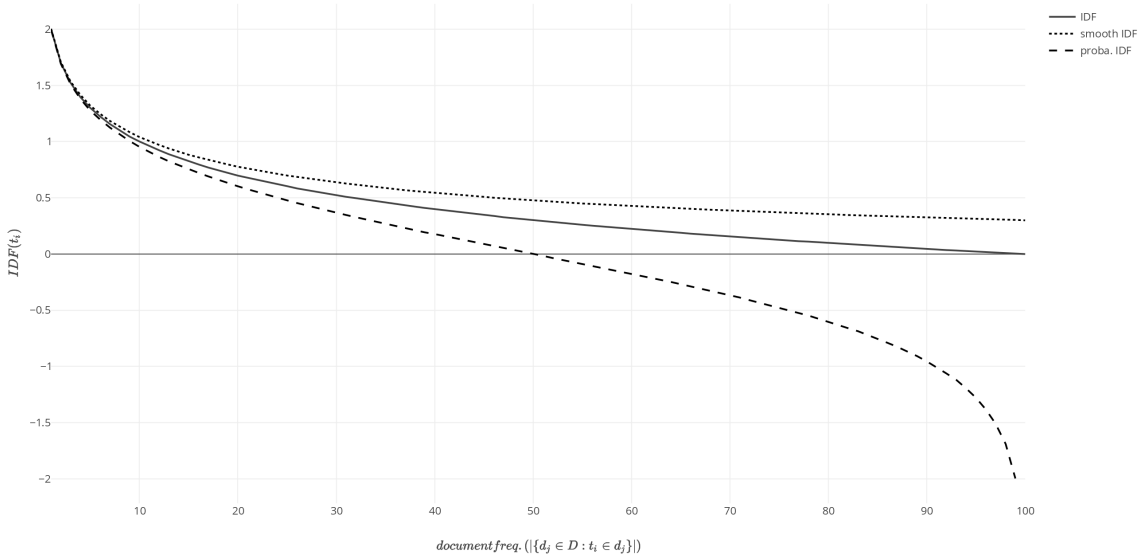


Figure 4.28: **The different IDF outputs.** A plot showing Inverse Document Frequency profiles against Document Frequency. This shows that the probabalistic IDF hilights words that are more important across all items, whilst the smooth IDF shows files which are more important individually. The general IDF (which is used) produces a result starting at 2 and tending to zero. This provides the best response and can easily be scaled between the range of [0,1] by dividing the output by 2. Source: [Mquantin, 2020]

Applying TF-IDF to chemical metrics

As the point of interest is in identifying species which are only important to a specific metric, and not across all, it is possible to adapt the TF-IDF equation. As the Term Frequency correspoinds to the number of times a value appears within the body of a document, the scaled metric value may be used. Next this is devided by the log of the Inverse Document Frequency. The Documnent Frequency can be given as the sum of values across all metrics. This makes the TF-IDF equation:

$$TF.IDF = metric\_value\ .\ \log(\frac{N_o\ documents}{\Sigma_\forall\ metric\_values}) \tag{4.14}$$

4.7.1 Metric Comparison

The aim of this section is to compare the efficiency of graph metrics against a list of traditional methods. To do this the use of a bivariate colourmap (Figure 4.29) is used. Each figure consists of a red hued image/heatmap representing the scaled values {0,1}:{white,red} for each of the individual columns. As each simulation contains thousands of species, only the top 10 species from each column/category are selected. These are then sorted by the average sum of their closeness, betweenness and page-rank values (blue column). Superimposed on this reds-only heatmap is a blue heatmap representing the average sum of the three metrics for comparison. Such a method allows for the comparison of individual values against a an approximation of species importance, by the sum of graph metrics. This allows us to partition the data into different categories.

- **Purple** - This value is high in both the individual category and the metric sum.
- **Red** - This value is high for the individual category but not the metric sum.
- **Blue** - This value is high for the metric sum but not the individual category.
- **White** - This value is low for all categories.

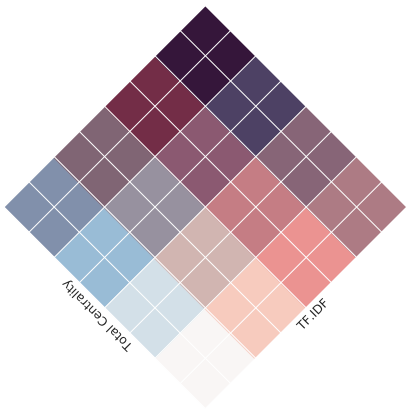


Figure 4.29: **The bivariate colourplot key.**

4.7.1.1 Individual Categories

Individual categories are split between traditional metrics and graph centrality metrics. To represent the importance of a species the following values may be extracted through the use of a simple box model:

- **Concentration** - This describes the abundance of a species within the atmosphere.

- **Net Flux** - This describes the rate of net (absolute) change of concentration over time for a species.

- **Absolute Flux** - Some species may have a large flux going through them (production and loss), resulting in a small net flux. This sums the production and loss fluxes.

- **Influence** - Influence is the total magnitude of an effect that changing a species concentration by 1% would have on other species within the network. Since the graph is generated using the Jacobian matrix, an alternative method for calculating this can be by calculating the total out-degree of a node.

The importance of a species is then compared through the use of three of the most common centrality metrics. These are:

- **Centrality** - This describes how easily information from one node can be disseminated to all other nodes.

- **Betweenness** - This describes the number of shortest paths (fastest fluxes/greatest influences) that are routed between nodes adjacent to our chosen node. Species with a high betweenness hold a brokering position, and can act as a bottleneck between different groups of chemistry.

- **PageRank** - PageRank looks at the flow in a system. It ranks nodes not only on the number of species it reacts with, but also the importance of the species it has reacted with

Finally the 'Metric Sum' is the sum of all the metric values scaled between 1 and zero (the mean).

4.7.2 •

4.7.3 what is important in all source location using reverse (adjoint) personalised page rank

Bibliography

(2019). LAPACK — Linear Algebra PACKage. `http://www.netlib.org/lapack/`. http://www.netlib.org/lapack/.

Barabási, A.-L. (2019). nature-150-cover.pdf. *Nature*, 575(7781). Accessed 20-01-2020.

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

Beel, J., Gipp, B., Langer, S., and Breitinger, C. (2016). Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4):305–338.

Bloss, C., Wagner, V., Jenkin, M. E., Volkamer, R., Bloss, W. J., Lee, J. D., Heard, D. E., Wirtz, K., Martin-Reviejo, M., Rea, G., Wenger, J. C., and Pilling, M. J. (2005). Development of a detailed chemical mechanism (mcmv3.1) for the atmospheric oxidation of aromatic hydrocarbons. *Atmospheric Chemistry and Physics*, 5(3):641–664.

Bloss, W. J., Lee, J. D., Bloss, C., Heard, D. E., Pilling, M. J., Wirtz, K., Martin-Reviejo, M., and Siese, M. (2004). Validation of the calibration of a laser-induced fluorescence instrument for the measurement of oh radicals in the atmosphere. *Atmospheric Chemistry and Physics*, 4(2):571–583.

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

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

Boudin, F. (2013). A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction.

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

Brasseur, G. and Jacob, D. (2017). *Modeling of Atmospheric Chemistry*. Cambridge University Press.

Broido, A. D. and Clauset, A. (2019). Scale-free networks are rare. *Nature communications*, 10(1):1017.

Cabello, R. (2019). three.js – JavaScript 3D library. `https://threejs.org/`. Accessed: 2020-1-25.

Cajal, S. R. (2020). Cortex drawings. *web*. Accessed: 2020-2-4.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703.

Cornell, L. (2020). Mark 1 Perceptron. `https://en.wikipedia.org/w/index.php?title=Perceptron&oldid=935763442`. Accessed: 2020-2-4.

de Solla Price, D. J. (1965). Networks of scientific papers. *Science*, 149(3683):510–515.

Derwent, R. G., Jenkin, M. E., Saunders, S. M., and Pilling, M. J. (1998). Photochemical ozone creation potentials for organic compounds in northwest Europe calculated with a master chemical mechanism. *Atmospheric environment*, 32(14):2429–2441.

Dillon, T. J., Tucceri, M. E., and Crowley, J. N. (2006). Laser induced fluorescence studies of iodine oxide chemistry part ii. the reactions of io with ch3o2, cf3o2 and o3. *Phys. Chem. Chem. Phys.*, 8:5185–5198.

Ebel, H., Mielsch, L.-I., and Bornholdt, S. (2002). Scale-free topology of e-mail networks. *Phys. Rev. E*, 66:035103.

Edsu and Ellis, D. (2019). etudier. https://github.com/wolfiex/etudier.

Ellis, D. (2019). Using TF-IDF to form descriptive chapter summaries via keyword extraction. `https://towardsdatascience.com/using-tf-idf-to-form-descriptive-chapter-summaries-via-keyword-extraction-4e6fd857d190`. Accessed: 2020-2-5.

Elshorbany, Y. F., Kleffmann, J., Hofzumahaus, A., Kurtenbach, R., Wiesen, P., Brauers, T., Bohn, B., Dorn, H.-P., Fuchs, H., Holland, F., Rohrer, F., Tillmann, R., Wegener, R., Wahner, A., Kanaya, Y., Yoshino, A., Nishida, S., Kajii, Y., Martinez, M., Kubistin, D., Harder, H., Lelieveld, J., Elste, T., Plass-Dülmer, C., Stange, G., Berresheim, H., and Schurath, U. (2012). HO x budgets during HOxComp: A case study of HO x chemistry under NO x -limited conditions. *Journal of geophysical research*, 117(D3).

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

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

Fujita, M., Inoue, H., and Terano, T. (2017). Searching promising researchers through network centrality measures of co-author networks of technical papers. In *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 615–618.

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

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.

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.

Google (2019). Google Scholar. `https://scholar.google.com/schhp?hl=en`. Accessed: 2020-1-25.

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

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, 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, 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.

Jenkin, M. E. and Hayman, G. D. (1999). Photochemical ozone creation potentials for oxygenated volatile organic compounds: sensitivity to variations in kinetic and mechanistic parameters. *Atmospheric environment*, 33(8):1275–1293.

Jenkin, M. E., Saunders, S. M., and Pilling, M. J. (1997). The tropospheric degradation of volatile organic compounds: a protocol for mechanism development. *Atmospheric Environment*, 31(1):81 – 104.

Jenkin, M. E., Saunders, S. M., Wagner, V., and Pilling, M. J. (2003). Protocol for the development of the master chemical mechanism, mcm v3 (part b): tropospheric degradation of aromatic volatile organic compounds. *Atmospheric Chemistry and Physics*, 3(1):181–193.

Jenkin, M. E., Young, J. C., and Rickard, A. R. (2015). The mcm v3.3.1 degradation scheme for isoprene. *Atmospheric Chemistry and Physics*, 15(20):11433–11459.

John Hay, Ben Lynch, D. S. (1960). MARK 1 PERCEPTRON OPERATORS' MANUAL. *Cornell Aeronautical Laboratory*.

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

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

Korke, R., Gatti, M. d. L., Lau, A. L. Y., Lim, J. W. E., Seow, T. K., Chung, M. C. M., and Hu, W.-S. (2004). Large scale gene expression profiling of metabolic shift of mammalian cells in culture. *Journal of biotechnology*, 107(1):1–17.

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

Kumar, R. and Upfal, E. (2000). The Web as a graph.

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

Ling, Z. H., Guo, H., Lam, S. H. M., Saunders, S. M., and Wang, T. (2014). Atmospheric photochemical reactivity and ozone production at two sites in hong kong: Application of a master chemical mechanism–photochemical box model. *Journal of Geophysical Research: Atmospheres*, 119(17):10567–10582.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

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.

Molontay, R. and Nagy, M. (2020). Twenty Years of Network Science: A Bibliographic and Co-authorship Network Analysis. *arXiv*.

Monastersky, R. and Van Noorden, R. (2019). 150 years of Nature: a data graphic charts our evolution. *Nature*, 575(7781):22–23.

Mquantin (2020). Accessed: 2020-2-5.

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

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5200–5205.

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

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, 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, 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.

poliaktiv (2011). Social Network Analysis: Theory and Applications.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C (2nd Ed.): The Art of Scientific Computing*. Cambridge University Press, USA.

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.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, pages 65–386.

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

Saunders, S. M., Jenkin, M. E., Derwent, R. G., and Pilling, M. J. (2003). Protocol for the development of the Master Chemical Mechanism, MCM v3 (Part A): tropospheric degradation of non-aromatic volatile organic compounds. *Atmospheric Chemistry and Physics*, 3(1):161–180.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269.

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.

Stubbins, A., Hubbard, V., Uher, G., Law, C. S., Upstill-Goddard, R. C., Aiken, G. R., and Mopper, K. (2008). Relating carbon monoxide photoproduction to dissolved organic matter functionality. *Environmental science & technology*, 42(9):3271–3276.

Turányi, T. and Tomlin, A. S. (2014). *Reduction of Reaction Mechanisms*. Springer Berlin Heidelberg, Berlin, Heidelberg.

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

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

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