Contents

Chapter 1

Chemical model diagnostics using graph theory and metrics.

"Entities should not be multiplied beyond necessity."

- William of Ockham, *Summa Logicae*

1.1 Introduction

In the previous chapters we have discussed visualisation and its role in bridging the gap between data and understanding. We have applied centrality metrics to a chemical network to tell us what species are of importance and experimented in getting various machine learning models to learn the chemical structure of the species involved in a mechanism. In this final research chapter we provide a (very) brief overview on mechanism reduction and propose two methods for simplifying the chemistry within a network.

Atmospheric Chemical Mechanism sizes have been increasing steadily in size over the last 10 years **??**. With the ability to automate mechanism construction, mechanisms with billions of species can be generated using a nubmer of pre-defined of rules - the mechanism protocol. Unfortunately as with large data, large mechanisms can pose a problem for the computation, visualisation and analysis of the chemistry. Having looked a methods to represent and query a mechanism, we now explore the different ways in which it may be simplified.

1.1.1 Reduction as a solution.

As mechanisms complexity has long been a problem, many methods of simplification and 'reduction' have been developed over the years. Although these are indeed useful, many reduced mechanism often require manual intervention and are usage/case specific. A generalisation of mechanism reduction is the elimination of species and reactions to produce a a smaller, more concise [1], mechanism whist retaining important properties or features of interest. There are many methods in the literature, the most common of which, are defined below.

1.1.1.1 Reaction Removal

The simplest method of reducing the number of items computed in a model, is to reduce the number of reactions. This eliviates the computational burden of calculating the rate of reaction each timestep. Classically this was done through the use of Rate and Production analaysis [ref previous chapter]. This allows for the visualisation of each reactions contribution to the rate of change of concentration of a a species of interest. In doing this we can may fillter redundant reactions that contribute less than a certain percentage[2] to the formation of our important species. Other more in depth methods include

[1] and thus manageable

[2] Typically 5%

the principle component analysis of the sensetivity (PCAS), where the most important parameters (the principal components) related to a simulation are selected. Here the objective parameters are those of our important species, and and the investigated parameters are the rate coefficients Vajda et al. [1985].

1.1.1.2 Species removal

Next we have species removal as a method to reduce a mechanism. This is useful not only because it reduced the size of the jacobian, but the removing or combining of species inherently reduces or simplifies the reactions within a mechanism. Whitehouse et al. [2004] states that using jacobian or sensetivity analysis methods proved 'capable' and 'efficient' in removing most redundant reactions and species from the MCM. Although there are many methods in which this may be done, all of these tend to partition the chemistry within three groups:

- **Important** - reactions or species directly related to the topic / outcome we are interested in
- **Needed** - reactions/species required by the important species, such that they may perform their desired function
- **Redundant** - those we may remove with little or no consequence to the final outcome of the model.

The interconnected, cascade nature of atmospheric chemistry results in the *important* species containing many dependencies. This means that many of these processes are iterative processes, where necessary species are added to the important species list on each itteration. This is then repeated until either a gap in the chemistry is reached, or more likely a mechanism of the desired size is obtained. There are several methods on how to approach this, some of which are outlined below.

Trial and Error

The simplest method is one of trial and error Turiinyi [1990] (Method 1). Here all the consuming reactions of a species are removed. If the resulting deviation between the full and the reduced mechanism is small, then the species may be removed. The downsides of this method is that it may be inefficient for large mechanisms, and only works on a per-species level - you cannot remove like groups.

Species Removal by Inspection of Rates

One of the first approaches to removing species was given by Frenklach in Oran and Boris [1991] with respect to combustion modelling. Here species whose reactions are much slower than the rate

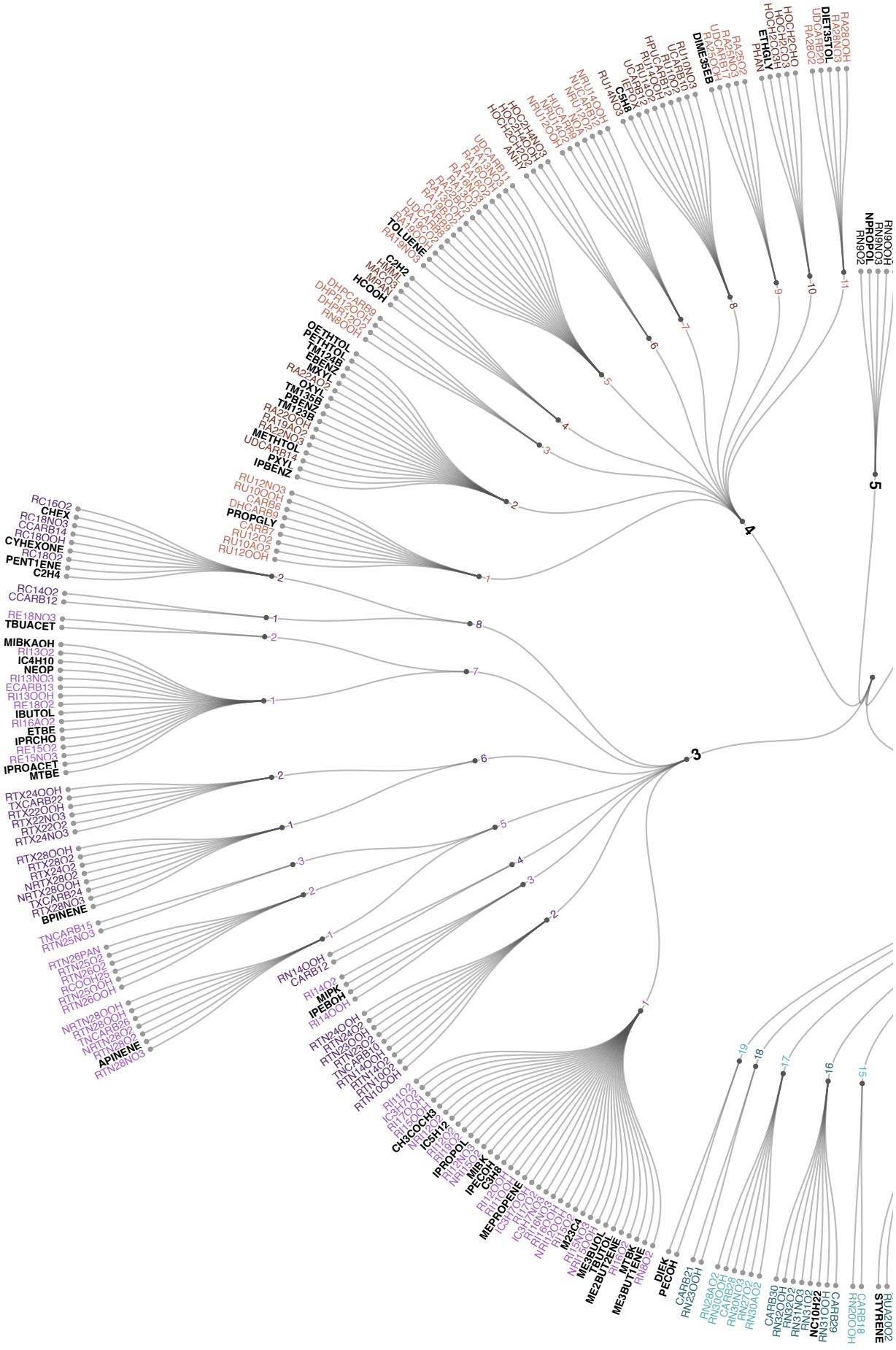
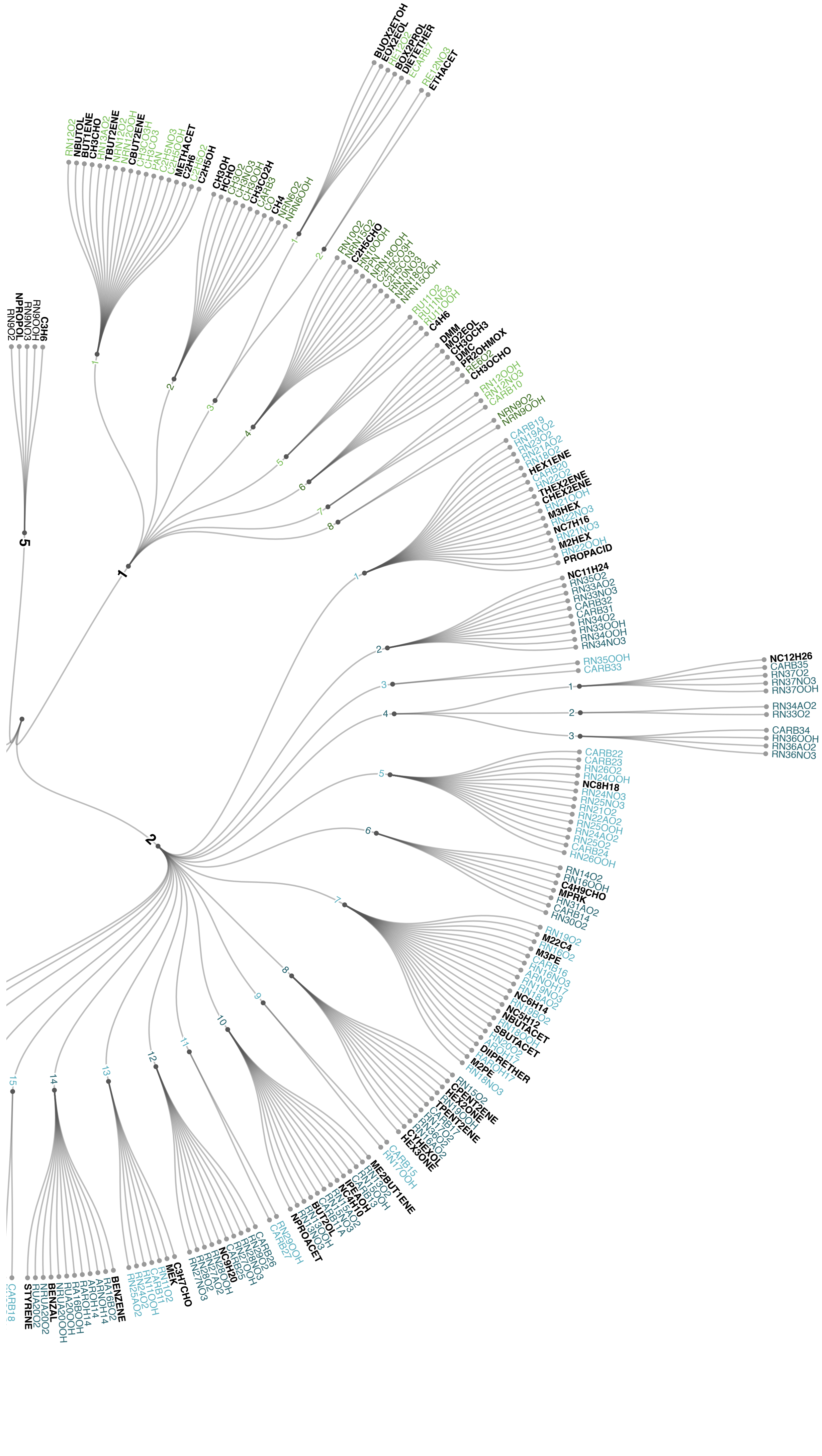


Figure 1.1: **A radial treemap showing the hierarchical clustering of the CRI mechanism.** The simulation results used are representative of the chemistry within London at Noon localtime and generated using DSMACC and the infomap algorithm.



5
C3H6
RN9OOH
RN9NO3
NPROPOL
RN9O2
1
RN12O2
NBUTOL
BUT1ENE
CH3CHO
RN13AO2
TBUT2ENE
NRN12O2
NRN12OOH
CBUT2ENE
CH3CO3H
CH3CO3
PAN
C2H5NO3
C2H5OOH
METHACET
C2H6
C2H5O2
C2H5OH
CH3OH
HCHO
CH3O2
CH3NO3
CH3OOH
CH3CO2H
CARB3
CO
CH4
NRN6O2
NRN6OOH
BUOX2ETOH
EOX2EOL
RE12O2
BOX2PROL
DIETETHER
ECARB7
RE12NO3
ETHACET
RN10O2
NRN15O2
C2H5CHO
RN10OOH
PPN
NRN18OOH
C2H5CO3H
C2H5CO3
RN10NO3
NRN18O2
NRN15OOH
RU11O2
RU11NO3
RU11OOH
C4H6
DMM
MO2EOL
CH3OCH3
DMC
PR2OHMOX
RE6O2
CH3OCHO
RN12OOH
RN12NO3
CARB10
NRN9O2
NRN9OOH
2
CARB19
RN19AO2
RN23O2
RN21AO2
RN18O2
HEX1ENE
CARB20
RN22O2
THEX2ENE
CHEX2ENE
RN21OOH
M3HEX
RN22NO3
NC7H16
RN21NO3
M2HEX
RN22OOH
PROPACID
NC11H24
RN35O2
RN33AO2
RN33NO3
CARB32
CARB31
RN34O2
RN33OOH
RN34OOH
RN34NO3
RN35OOH
CARB33
NC12H26
CARB35
RN37O2
RN37NO3
RN37OOH
RN34AO2
RN33O2
CARB34
RN36OOH
RN36AO2
RN36NO3
CARB22
CARB23
RN26O2
RN24OOH
NC8H18
RN24NO3
RN25NO3
RN21O2
RN22AO2
RN25OOH
RN24AO2
RN25O2
CARB24
RN26OOH
RN14O2
RN16OOH
C4H9CHO
MPRK
RN31AO2
CARB14
RN30O2
RN19O2
M22C4
RN16O2
M3PE
CARB16
RN16NO3
ARNOH17
RN19NO3
RN18AO2
NC6H14
RN19BO2
NC5H12
NBUTACET
RN18OOH
SBUTACET
RN20O2
AROH17
DIIPRETHER
RAROH17
M2PE
RN18NO3
RN15O2
CPENT2ENE
HEX2ONE
RN19OOH
TPENT2ENE
CARB17
RN17O2
RN36O2
RN16AO2
CYHEXOL
HEX3ONE
CARB15
RN17OOH
ME2BUT1ENE
RN13O2
RN15OOH
CARB13
IPEAOH
NC4H10
RN15AO2
RN15NO3
CARB11A
BUT2OL
RN13OOH
RN13NO3
NPROACET
RN29OOH
CARB27
CARB26
RN29O2
RN28NO3
RN27OOH
CARB25
NC9H20
RN28OOH
RN27AO2
RN28O2
RN27NO3
C3H7CHO
MEK
RN11O2
CARB11
RN11OOH
RN24O2
RN25AO2
BENZENE
RA16BO2
ARNOH14
AROH14
RAROH14
RA16BOOH
RUA20OOH
NRUA20OOH
BENZAL
NRUA20O2
RUA20O2
STYRENE
CARB18

determining steps of a mechanism are marked as redundant and removed.

Jacobian Connectivity Method

The log-normalised Jacobian matrix can be used to determine the change in production of a species due to a 1% change in the concentration of any other species. In squaring the effect of this for all improtant species, we get a metric depicting how the change in a certain species affects the concentrations on all important species, Equation 1.1, where $(y_i/f_i)(\partial f_i/\partial y_i)$ is element i of the normalised Jacobian . Through an iterative process we can identify redundant species, of a low contribution to our important species, and remove them. This is known as the Connectivity Method Turányi [1990].

$$B_i = \sum_j ((y_i/f_i)(\partial f_i/\partial y_i))^2 \tag{1.1}$$

1.1.2 Lumping

Rather than removing species or reactions from a mechanism we may combine them to form a new composite species. This is species lumping. To do this we must first consider how we determine species that are to be joined together, and then how their grouped reactions will contribute to every other species it reacts with. Some of the more general types of lumping styles are outlined below.

1.1.3 Chemical Lumping

Mechanisms follow protocols in their generation. This produces reaction styles that many like-structured species follow in their degregation. In determining such classes we may be able to generalise like-species reactions and group them together as one. Taking the CRI mechanism as an example, this has taken the Ozone production capability as a feature of interest. The rato of CC and CH bonds are used to determine a species oxidation possibility ... An example of this are species such as CARB9 ... some examples and what the names mean

As this type of lumping has already been performed on our starting mechanism, we shall not be applying it any further.

1.1.3.1 Linear

1.1.3.2 Lifetime

1.1.3.3 Quasi Steady State Approximation (QSSA)

QSSA works on the axiom that the the flux through a species is 0 - Use louise whitehouses thesis here (better description that the analysis to kinetic reactions book)

1.1.4 Graph parallels.

Although there are many graph based methods that exist within the reduction realm, most of these concentrate on the generation of skeletal methods through the building of a directed tree (subcategory of graphs from source to target) - LIST of refs and sentence of all skeletal methods. Path flux analysis (Sun et al 2010)

Instead we may find ourselves applying graph theory to solve other reduction methods. For instance we can trace back influence through connecting edges using dijkstras shortest path algorithm (CH2 ref) - analogous to the connectivity method, or a leave one out approach combined with pagerank to access the effects of removing a node.

We can use the graph structure to analyse changes of reactions or relationships between species. This can provide an alternative representation and method to access such data. Additionally we may use graph clustering techniques to locate groups of highly connected, fast reacting/strongly related species. This has applications in both understanding the data, but more importantly chemical lumping. In creating a graph from the mechanism, we not only encode information about the chemical structure, but also the rate of reaction in the graph. In grouping species by high numbers of reactions between them with fast fluxes we can take a QSSA style approach to reduction, and assume that since the rate of reaction between them is much much faster than those outside a cluster, they may be grouped together. This will be explored in PART II [ref link].

1.2 Methodology: Part I

In this chapter we are not interested in reducing a mechanism for a certain case study or environment. Instead we look at tools capable of simplifying the hard coded maths behind the reactions that exist within the atmosphere. Since the key chemical drivers differ from location to location we make use of

our ability to run a range of different simulations which provide an overview of the entirety of the input space (intial concentrations). Although this may not replicate and real-word (physical) scenarios, it tests the robustness of the mathematics behind the mechanism and in using this data to predict like properties based on the physical equations which describe the system. We work on the assumption that these have been correctly fitted and tested against experimental data, and that through reducing the mechanism based purely on its mathematical response to different conditions, its ability to predict atmospheric science will be preserved.

The model setup shall be defined below.

1.2.1 The Mechanism

The mechanism we wish to use as a baseline is the Common Representative Intermediates (CRI) Mechanism of version 2.265 ,Jenkin [2019]. This is an update to the CRI v2.0, with the purpose of updating the chemistry to better represent that of MCM version 3.3.1 (i.e. the inclusion of explicit Isoprene chemistry). The CRI v2.265 mechanism has been developed much in the same way as its precursor, and is centred around describing the ozone formation from Volatile organic compounds (VOCs) in the troposphere. The main assumption behind the lumping in the generation of this mechanism is that 'the potential for ozone formation from a given volatile organic compound (VOC) is related to the number of reactive (i.e., CC and CH) bonds it contains',Jenkin et al. [2008]. Reductions are made on a compound-by-compound basis and compared to the MCM using a series of 5 day box-model simulations.

The CRI v2.265 mechanism contains 422 species and 1261 reactions. This is still almost double of those within the global GEOS-Chem model. With explicit manual reduction being used to reduce the MCM by YY percent, we wish to apply a series of automated reduction techniques and compare them to the baseline mechanism. We chose the 2.265 version of CRI, since it posesses both a potential for further reduction (CRI v2.0 was reduced a further 5 times); there are a sufficient number of items to make it a non-trivial example, but mainly as its relative size is just within our ability to visually inspect the output of each species.

1.2.2 The Box-Model

The box model used shall be an adapted version of the Dynamically Simple Model of Atmospheric Chemical Complexity (DSMACC) [ref doi, ref DSMACC]. This has had several changes which allow

for multiple parallel runs, easy extraction of rates, fluxes and the jacobian matrix as well as a simple ncurses interface for loading and parsing new files.

The DSMACC model works by using the Kinetic Pre Processor (KPP) [REF] to generate Fortran code, which can then be used to integrate the provided mechanism. As there were some issues presented with this a pre-pre parser code was used on the mechanism before running KPP, and a post parser on some of the files to provide the desired output.

1.2.3 Model Inputs

As we do not wish to constrain results to a specific case example, we provide a range of inputs with the aim of covering all possible concentration combinations. As a means of determining all non-lumped species within the CRI mechanism, we only used those which appear in both the MCM and CRI. This also ensures that should it be desired at a later date, any further reduced mechanisms can directly be compared with the results of the full MCM.

In terms of sampling there are several types that may be used ,Mckay et al. [2000]. The main or most common is Random, or Monte Carlo, sampling. The problem with this is that its pseudo random nature may often results in regions of high density and some of low, Figure **??**. In attempting to sample the entirety of our input space we use a method based on the principles of a Latin square. A latin square is a square containing n items, arranged in such a way that they only appear once in each row and column, much akin to a sudoku puzzle ,lsq [2008]. A latin hypercube is a generalisation of this allowing for an alternative subspace sampling to the monte carlo method which expands the probabilistic stratified-style sampling of a latin square into n-dimentional space. This presents better covarage of our input space Figure **??** and will be used with the following limits, TAABLE, to gen our sample simulations.

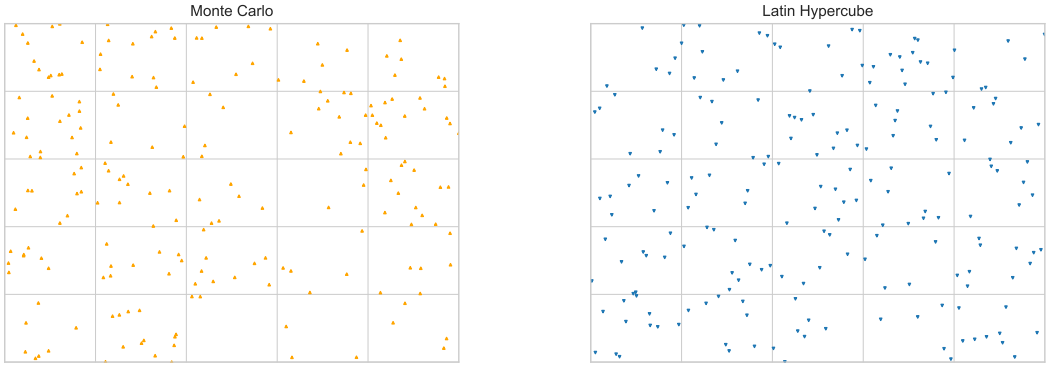


Figure 1.2: A comparison of the distribution of 250 sampled points using a) Monte Carlo and b) Latin Hypercube sampling

$$\text{concentration} \begin{cases} min = 10^{-8}\ max = 10^{-13}, & \textbf{if} NO, NO_2, O_3 \\ min = 10^{-8}\ max = 10^{-13}, & \text{otherwise} \end{cases} \tag{1.2}$$

1.2.4 Reduction through Lifetime

1.2.4.1 Calculating the lifetime

Within models a species lifetime is regarded as the time taken for its concentration to halve [ref]. This works on the assumption that the species is not produced, and that rate coefficients and other constants remain constant. For a first order decay of sample Equation 1.3, we can represent the decay using Equation 1.2.4.1, showing that the half life is independent of initial concentration.

$$A \rightarrow {}^{k} B \tag{1.3}$$

$$s(t) = a_0 \exp(-kt) \frac{a(t)}{a_0} = \exp(-kt)$$

linearised this gives

$$\ln(\frac{a(t)}{a_0}) = -kt$$

after $\tau_{1/2}$ the concentration is equal to $a_0/2$ of initial rate a_0, which gives

$$\ln(\frac{\frac{a_0}{2}}{a_0}) = \ln(\frac{1}{2}) = \ln(2^{-1}) = -\ln 2 = k\tau_{1/2}$$

$$\tau_{\frac{1}{2}} = \frac{\ln 2}{k} \tag{1.4}$$

In species of the first order only, this may simplified to

$$a(t) = a_0 \exp(t \sum_j k_j)$$

and therefore the half life may be written as the reciprocal sum of rate coefficients:

$$\tau_A = 1/\sum_j k_j \tag{1.5}$$

and is how lifetime is calculated for photochemical species [ref! pillin and seakins]. An alternative method for half life calculation may be obtained using the diagonal (self reference) of a Jacobian matrix ,Turanyi and Tomlin [2015]:

$$\tau_1 = -\frac{1}{J_{ii}} \tag{1.6}$$

This value will usually be negative unless a species does not contain a consuming reaction, then it will be zero.

The xxxxx method of reduction consists of the isolation of species with similar lifetimes and reactions as a means of lumping. In doing so the ... etc

1.2.5 Comparing Magnitude and Direction

Since the photolysis reactions in a model change the resultant rates, and thus flux of a species depending on the azimuthal angle related to the time of day, we not only want to compare species with the same magnitude, they also need to match the profile as they change. To do this we may represent all pariwise species matches on a latent space representing the size and angle between their temporal vectors. This is done through using the euclidean distance on the x axis, and cosine distance y on the y.

1.2.5.1 Euclidian distance

This is the simplest method of vector comparison and works by calculating the distance between all points in two vectors. For the vectors

$$v1 = [a, b, c, \ldots n]$$

$$v2 = [i, j, k, \dots z] \tag{1.7}$$

This can be done using pythoagoras' theorem in Equation 1.8:

$$e_{dist} = \sqrt{(a-i)^2 + (b-j)^2 + (c-k)^2 + \dots + (n-z)^2} \tag{1.8}$$

This transformation converts the straight line distance between each vector into metric space, allowing us to represent the difference in their magnitudes as a single scalar. Unfortunately as this requires the difference between all permutations of rows, it cannot be done as a single operation, but as multiple.

APPLICA"tiion

1.2.5.2 Cosine Distance

Similarly if we wish to calculate the angle between two vectors we may use the cosine difference. In starting with the definition of the dot product

$$v1 \cdot v2 = \|v1\|\|v2\| \cos\theta$$

this may be arranged

$$\cos\theta = \frac{v1 \cdot v2}{\|v1\|\|v2\|} \tag{1.9}$$

Since this does not work for the triangle ? inequality, we need to normalise each vector before calculating the cosine distance. The merits of this come from ... which makes its application comparing the similarity between texts or documents of different sizes very popular (REF!).

COMPARE force graph of cosine differences and force graph of euclidean distances. Colour ones close to eachother.

1.3 Results: Part I

In order to get a representation of the mechanism we run 300 randomly initiated scenarios, with the experimental design capable of accepting more data at a later point. We then extract the lifetimes from the diagonal of the jacobian, and convert them into vectors representing the duration of all 3 days for each simulation. These vectors are then compared against eachother to find species of similar

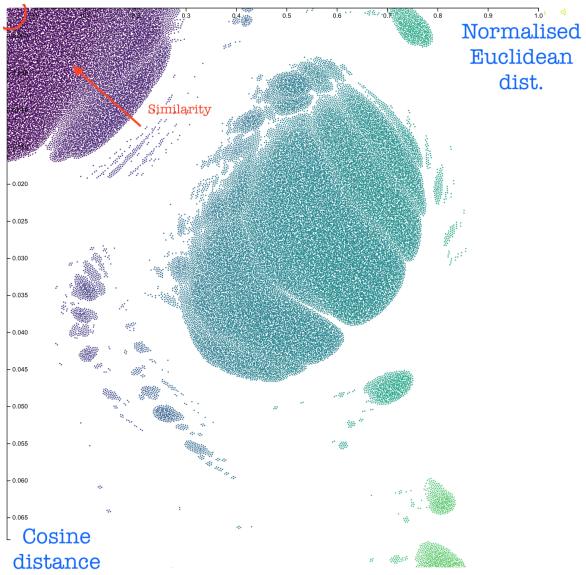
lifetimes, which follow a like response to the diurnal cycle. This is done through the use of euclidian and cosine similarities.

1.3.1 Temporal Lifetime Vector Comparison

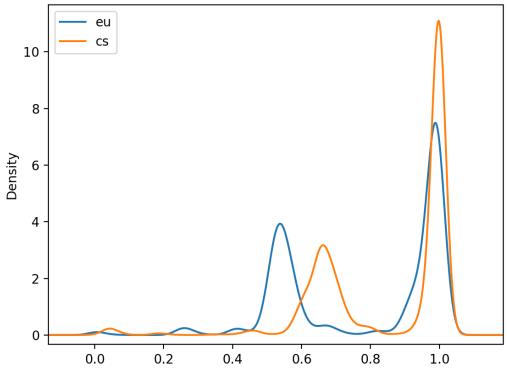
First we wish to explore the results of a single simulation. This will help establish the thresholds and methodology that is to be used in future examples. As we do not wish to lump inorganics, we remove these from the program and generate all the pairwise interactions between species. We compute the euclidean and cosine distances for all remaining reaction pairs (88410 pairs) across the entirety of the single calculation. These are then plotted on separate axis, and coloured using the geometric mean, Figure **??**.

Since there are n^2 different combinations, this process can be lengthy to compute. Moreover the number of nodes present makes it difficult to visually determine which couplets to lump together, especially when they are overlaying eachother, as in Figure **??**. To overcome this we apply a force graph simulation to each node. Here a strong force pulls them towards their true location, whist a collision/repulsive force ensures that there exists no overlap between nodes. Although some information loss may be incurred this way, it does enable us to interactively determine which pairs belong to which points.

Plotting the data this way provides an approximate view of how lifetimes change within the system. Figure **??** shows the distribution of values across both metrics. Here we see that although there is some difference the cosine difference can be used as an indicator for exploring further. This is useful since the cosine difference is a matrix operation and can produce near instant results in the form of the output relationship matrix. The euclidean distance however must be computed for each coupling individually as it requires taking the difference between the temporal lifetime arrays of each species. This means that rather than computing the euclidean distance for all permutaions of species, we can compute only those whose cosine distances suggest potentially promising results.

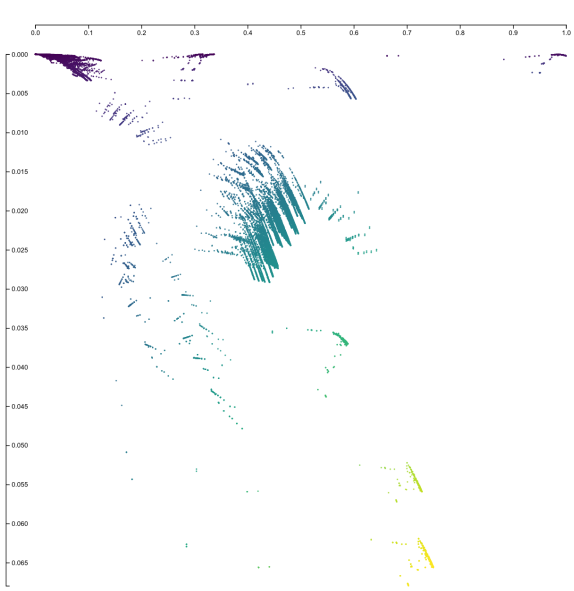


(a) Interactive, non overlapping plot of normalised euclidian distance across the x axis against the cosine distance on the y. The colouring is the geometric mean between them.

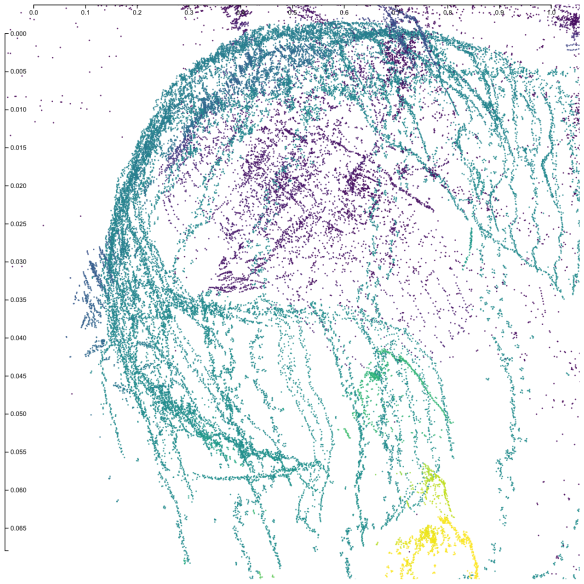


(b) Gaussian Kernel Density Estimate plot showing the distributions present for each distance metric

Figure 1.4: Showing the evolution from the original overlaid locations, Figure **??** to the slightly more accessible (interactively) Figure **??**



(a) Original



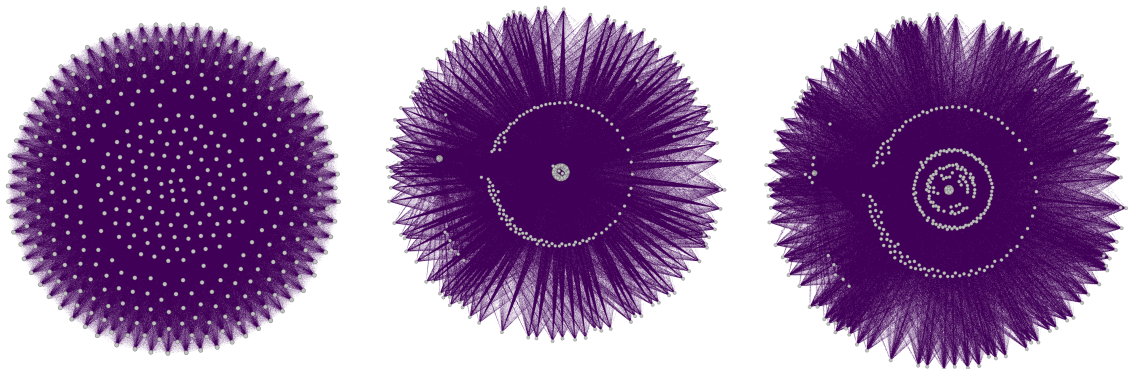
(b) First collision detection timestep, before attractive forces have fully kicked in.

Figure 1.3: Showing the evolution from the original overlaid locations, Figure **??** to the slightly more accessible (interactively) Figure **??**

1.3.2 Viewing the similarity graph

One method to simplify this is to convert to the pairwise interaction list into a fully connected graph. Here we have individual species, pulled together by the similarity (geometric mean) between each two nodes. Unfortunately in being a complete graph of 421 species and 88410 links, this is quite difficult to visualise without forming a *hairball* [fig a]. As a means of filtering the results we trim the weakest links in sequence. This produces rings of densely connected species based on lifetime thresholds which corresponds to the different types of chemsitry species undergo

Need to find out what the abbreviated names in cri stand for!. The sizes and species within each ring show similararities to the distance metric plot as would be expected. This provides an alternative representation where graph community detection algorithms such as MCL [ref] can be used to isolate chemistry with different levels of lifetime relationships between them. Unfortunately, as we are interested species with near identical lifetimes, this does not hold a great deal of merit in locating these.



(a) Initial connected graph

(b) Initial graph with trimming half the edges

(c) Heavy trimming.

Figure 1.5: A froce graph with progressive trimming of weak, poorly related, edges.

1.3.2.1 Using MCMC to extract groups

Need to see what Cri names mean to better explain these.

Bibliography

(2008). *Latin Square Designs*, pages 297–297. Springer New York, New York, NY.

Jenkin, M. (2019). Http://Cri.York.Ac.Uk . Online.

Jenkin, M., Watson, L., Utembe, S., and Shallcross, D. (2008). A common representative intermediates (cri) mechanism for voc degradation. part 1: Gas phase mechanism development. *Atmospheric Environment*, 42(31):7185 – 7195.

Mckay, M. D., Beckman, R. J., and Conover, W. J. (2000). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61.

Oran, E. and Boris, J. (1991). Numerical approaches to combustion modeling. progress in astronautics and aeronautics. vol. 135. *U.S. Department of Energy Office of Scientific and Technical Information.*

Turányi, T. (1990). Sensitivity Analysis Of Complex Kinetic Systems. Tools And Applications. *Journal of mathematical chemistry*, 5(3):203–248.

Turanyi, T. and Tomlin, A. (2015). *Analysis Of Kinetic Reaction Mechanisms*. Springer.

Turiinyi, T. (1990). Reductton Large Reactton Mechantsms. *New journal of chemistry = Nouveau journal de chimie*, 14:795–gO3.

Vajda, S., Valko, P., and Turainyi, T. (1985). Principal component analysis of kinetic models. *International Journal of Chemical Kinetics*, 17(1):55–81.

Whitehouse, L. E., Tomlin, A. S., and Pilling, M. J. (2004). Systematic reduction of complex tropospheric chemical mechanisms, part i: Sensitivity and time-scale analyses. *Atmospheric Chemistry and Physics*, 4(7):2025–2056.