Contents

1 Chemical mechanism stratification and analysis using ML and graph clustering. 1

1.1 Introduction . . . 4

1.2 Mechanism Reduction . . . 4

1.2.1 Reaction Removal . . . 5

1.2.2 Species removal . . . 5

1.2.3 Lumping . . . 6

1.2.3.1 Chemical Lumping . . . 6

1.2.3.2 Linear . . . 7

1.2.3.3 Lifetime . . . 7

1.2.3.4 Quasi Steady State Approximation (QSSA) . . . 7

1.3 Data Setup . . . 7

1.3.1 The Mechanism . . . 7

1.3.2 The Box-Model . . . 8

1.3.3 Model Inputs . . . 8

1.4 Graph based reduction . . . 8

1.4.1 Graph parallels. . . . 9

1.4.2 Types of Graph Clustering . . . 9

1.4.3 Walk/Flow Based Clustering . . . 10

1.4.4 Infomap for graphical clustering . . . 11

1.4.4.1 Species type and clustering . . . 14

1.4.4.2 Inter and Intra links . . . 15

1.4.4.3 Number of clusters . . . 18

1.5 Reduction through Lifetime . . . 19

1.5.0.1 Calculating the lifetime . . . 19

1.5.1 Comparing Magnitude and Direction . . . 21

1.5.1.1 Euclidian distance . . . 21

1.5.1.2 Cosine Distance . . . 21

1.5.2 Temporal Lifetime Vector Comparison . . . 22

1.6 Results . . . 23

1.6.1 The co-grouping network . . . 24

1.6.2 Comparing daytime and nightime groups . . . 25

1.6.3 Determining cluster suitabiltiy . . . 27

Chapter 1

Chemical mechanism stratification and analysis using ML and graph clustering.

"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 brief overview on mechanism reduction and propose two methods for simplifying the chemistry within a network.

Science often deals with the problem of understanding complexity. Often this may be accomplished through organisation and partitioning, for example the learning of a new skill, or the parallelisation of a large mathematical problme. In cases where such methods fail, we are forced to 'disregard' complexity. It is common to aproximate an atom as a sphere or the value π as 3 with little consequence to the overall result of a calculation. The process of lumping has long been used replace a complex, changing process (e.g Quantum Mechanics or Boundary Layer Fluid Dynamics) with a simpler constant process, [Mahajan, 2008]. In such cases an approximate analysis may be far more useful than a lengthy exact solution, or none at all.

Similar problems of complexity can also be seen within the chemistry of the atmosphere. An example is seen within the Master Chemical Mechanism[1] (MCM), [**?**], containing 1228 RO_2 reactions. If written explicitely all $RO_2 - RO_2$ interactions would result in the addition of 1507984 reactions. Instead the MCM overcomes this problem by createing an RO_2 pool, with which all RO_2 species react. This results in a mechanism which produces similar results, but only contains 0.000814 of the total possible $RO_2 - RO_2$ reactions.

However even with such simplifications atmospheric chemical mechanisms have been increasing in size over the last 10 years, **??**REF. With the ability to automate their construction, mechanisms with species numbers of the millions become possible. Although the existence of more-explicit mechanisms may improve the quality of science produced, they can cause problems for efficient compuation, diagnosis and anlysis. This chapter shall look at two methods in which we may simplify a mechanism by grouping similar species together.

1.2 Mechanism Reduction

As discussed, the first step to simplifying a complex task involves the partitioning data into categories. For a mechanism we begin by looking at the reaction or species which are related to the area that is

[1] Version 3.3.1 .

being researched. Items are partitioned into important, needed and redundant categories (described below).

- **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.

1.2.1 Reaction Removal

Since atmospheric chemical mechanism form a numerically stiff system, a reduction in the number of reactions within a mechanism leads to a reduction in the computational burden experienced by a model each iteration forwards in time. Classically the identification of important reactions may be accomplished through the use or rate of production and loss analysis (SEC REF). This allows us to filter reactions contributing less than 5% to the formation of any species we are interested in. Other methods using principle component analysis of the sensetivity of species (PCAS) also exists and are discussed in Vajda et al. [1985].

1.2.2 Species removal

Similar to reaction removal, species removal 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. This method also has added benefit of reducing the size of the jacobian matrix used to propagate the chemical system forwards. For large systems which do not use a sparse framework, storing an $n * *2$ matrix in memory can prove difficult.

There are many methods of species reduction that are possible. The simplest of these is through the use of trial and error[2] [Turiinyi, 1990] (Method 1). Here the consuming reactions for a species are removed, and if the resulting deviation in results between the full and reduced mechanism is small, ther results are kept. The main downside to this is that it only works on a per-species level, which may be very resource consuming for large mechanisms.

Alternatively it is also possible to remove species whose reactions are much slower than the rate determining steps of a mechanism, [Oran and Boris, 1991]. Although this method is useful in the field of combustion modelling, this has most likely been done for atmospheric chemical mechanisms.

[2] A tried an tested method for scientific discovery.

Finally the use of jacobian sensetivity analysis has intensively been used in the determination of which species may be removed from a mechanism. Whitehouse et al. [2004] states this to be a 'capable' and 'efficient' method for removing most redundant reactions and species from the MCM. Use of a log-normalised Jacobian to determine which species can be removed is found in the connectivity method **??**. Here the influence in changing a 1% change in a species concentration has on the concentration of important species can be determined by

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

where $(y_i/f_i)(\partial f_i/\partial y_i)$ is element i of the normalised Jacobian. Through an iterative process species with a low contribution to our important species can be dound and removed.

1.2.3 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.

1.2.3.1 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.

1.2.3.2 Linear

1.2.3.3 Lifetime

1.2.3.4 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)

1.3 Data Setup

Unlike manual reduction, this chapter does not concern itself with the intricacies of the chemistry behind a mechanism. Instead serach for an automated method of simplifying the mathematical structure behind a mechanism whist preserving the quality of science it represents. Although this method of analysis may not directly replicate the real-world, it can provide an acurate test of the robustness of a mechanism and the equations within it. I work on the assumption that the equations describing each reaction are representative of experimental results, and in simplifying these, their usefulness in modelling the real data is preserved. This section describes the experimental setup for the experiment.

1.3.1 The Mechanism

The mechanism used is the Common Representative Intermediates (CRI) Mechanism v2.2 ,Jenkin [2019]. This is an already reduced version of the MCM,where species are grouped based on their ozone formation potential - i.e. the C−C and C−H ratio of bonds. Reductions have been made on a compound-by-compound basis and compared to the MCM using a series of 5 day box-model simulations,Jenkin et al. [2008].

Why further simplify the CRI network

CRI v2.2 is a mechanism containing 422 species (1261 reactions). Although this is significantly smaller than the full MCM, it may still prove problematic if used within a global model - for comparison the GEOS-Chem standard chemistry is approcimately half the size of this. Additionally the 2.0 was reduced a further 5 times from its base size. Since this has yet to be done for v2.2, it would be an interesting test to see how far it may be reduced using new and untried methods, whilst still being able to manually inspect simulation results.

1.3.2 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.

1.3.3 Model Inputs

The aim of this experiment is not to replicate a specific case study or senario. Instead we extract all non-lumped species which appear in both CRI and the MCM and provide an assortment of initial condition concentrations to cover the entriety of the input space.

To select the inital conditions there exist several sampling styles Mckay et al. [2000]. The most common style is the random or 'Monte Carlo' approach, however this does not guarantee a homogenious distribution of points. A lattice or grid approach is also possible, but that can result in a large number of sample points to produce an complete distribution of the input space. To overcome this a latin hypercube can be used. This is a generalisation of the latin square - a square matrix containing n items, arranged in such a way that they only appear once in each row and column (akin to a sudoku puzzle) lsq [2008]. This presents better covarage of the input space, which shall be setup as follows:

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

This process is used to create the inital conditions files for X species and 300 simulations.

1.4 Graph based reduction

It has been shown that the graph-based representation of the atmospheric chemical network proves useful in both the visual and mathematical analysis of simulation results (**??**). It therefore follows that the network representation of mechanism may also have its uses in the simplification, and thus

reduction, of chemical complexity. This section will outline the basic methods of modularity (or clustering) detection with the graph framework, the different methods in which this may be done and eventually apply it to a case example representative of the chemistry within the London environment.

1.4.1 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].

1.4.2 Types of Graph Clustering

Unlike vector clustering algorithms (such as DBSCAN, UMAP and K-means, as discussed in SEV**??**), graph clustering metrics do not rely on the spatial orientation of the data to determine groups or 'clusters'. Instead these may partition the network into segments, group nodes by structual equivalence or explore the 'flow' dynamics of the network.

Algorithms such as Label Propagation [Raghavan et al., 2007] and spinglass [Newman and Girvan, 2004] work by randomly assigning nodes with a property or label. This property is then transferred to its neighbours. Other algorithms such as the nested block model can decompose a graph into clusters

of like properties, [Fortunato, 2010]. These are often grouped in the form of topological equivalence which can be either:

- *structual equivalence* - vertices are similar if they have like neighbours, [Zhou, 2003].
- *regular eqivalence* - retrieves nodes with similar connection patterns (e.g. parent - child node hierarchicl structures), [Everett and Borgatti, 1994].

This works in a similar way to an autoencoder (ref auto **??**), where topological similarities are used to simplify (or encode) the network structure, in a way which it may be again decoded.

Finally there exist a set of 'flow' based models which use the network dynamics to determine the modularity of a network. These are discussed below.

1.4.3 Walk/Flow Based Clustering

As temporal networks result in a change (magnitude, or type) of relationships between items. Such changes in the network dynamics are encoded within the edges of a graph. To account for this, the primary function of random walk or 'flow' algorithms is to capture the the changes between the real-world systems represented by the network.

In (SECTION SILHO) the silhouette coefficient was discussed. This compares the vector position of clusters with regards to the distance of data points between them. Translating this to the graph framework, topological (graph) clustering defines a cluster, or module, as a region with a greater inter-cluster degree or density[3] compared to their intra-cluster density[4]. This results in a system, that if sorted by group, has more links between elements of the same group than with those in other groups - this can be seen within the sorted adjacency matrix in (Xhaapter 1 REF).

Since flow based methods are more interested in the network dynamics, than structure, the number of links or density is replaces with the time a random 'walker' spends 'trapped' between a set of nodes. A real-wold analogy would be to view the flow of water in a slowly filling river, Figure 1.1. Here a walker (or water molecle) traverses the entirety of the river/graph network, occasionally getting trapped between a set of nodes. Here although the water is still moving, it ends up spending more time going back and fort between a set of nodes, than exploring the rest of the network. It is these regions of stalled progress that form our network clusters.

[3]The number of links or edges between items in the same group.

[4]The number of edges to other clusters

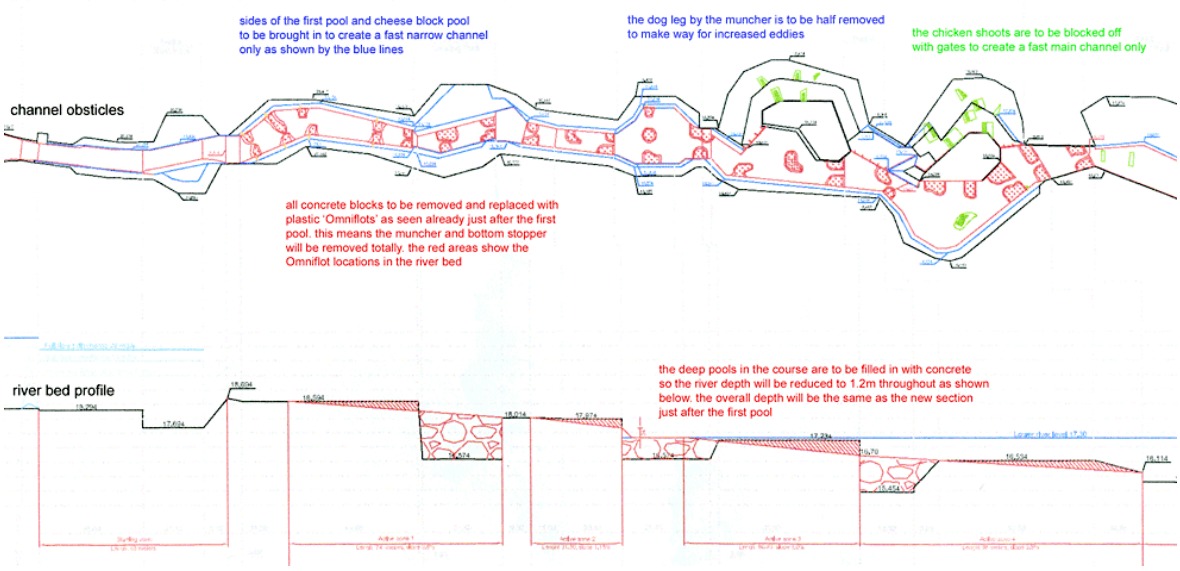


Figure 1.1: **The proposed plans for the change of the UK National Watersports Centre Whitewater Course (Holme Pierrepont).** Walk based clustering is analogous to the movement of a river. Clusters (or modules) are identified as areas where the 'flow' becomes trapped, much like water in the pools immediately following a hydraulic jump. Source: Cornes [2008]

The Loudvain clustering algorithm is one of the most popular of the clustering algorithms due to its algorithmic and qualatative robustness, [Blondel et al., 2008; Lu et al., 2015]. On the simplest level this works by maximising the modularity for each configuration. Modularity is a value between positive and negative unity which measures the density of edge between inter and intra communities and compares it to an equivalent random network. The Loudvain is a hierarchical clustering algorith, this means that after each iteration all nodes which belong to the same cluster are consolidated to form a new 'grouped' item. Inter-cluster links are converted into self-links, and intra-cluster links are updated accordingly [REF INCLUDE LAYERS OF hierarchi VCRI'

Similar to the Loudvain algorithm is the Rosvall and Bergstrom [2008]'s Infomap. Here each node within the network is assigned its own module. These are then perturbed to neighbouring nodes should such a mode lead to a decrease in the map equation (a flow-based method which operates on system dynamics rather than structure - [Rosvall et al., 2009]). The process is repeated until no further reductions are possible.

TWO LEVEL

MULITLEVERL

1.4.4 Infomap for graphical clustering

Although there are a wide range of availiable clustering technques, there are several reasons that the infomap method works well. The main two criterions in selecting an algorithm for grouping

atmospheric reactions are that the algorithm can deal with a directed network (chemistry is directional) and that it needs to be able to handle temopral data (since concentration and photolysis conditions change with time, this plays an important role within the system). The infomap algorithm implements a directed approach, improving on the core loudvain core algorithm it is built upon. In addition to this it is able to implement a multi-level clustering approach which has been shown to capture nodel-layer interaction in temporal networks,[Aslak et al., 2018].

Using the intial conditions for London from the last chapter (**??**), a spun up simulation run with the CRI v2.2 mechanism was run. Since this does not contain $C_5H_{11}CHO$, MVK, MACR or Limonene, these species are omitted from the initialisation. Following a spinup to steady state, a graph is generated for noon after 1 day of an unconstrained run. The infomap algorithm is then applied to the generated graph.

The coursest level of clustering is shown in Figure 1.2. Here nodes are coloured by their cluster, and approximate polygon hulls surround the nodes closest to the median cluster centre. Much like the findings in (CHAPTER STRT), it is seen that different sections of the graph network repesent different types of chemistry - for example hull 4 contains aromatic species, hull 2 contains the products of linear alkanes and hull 3 contains the terpines.

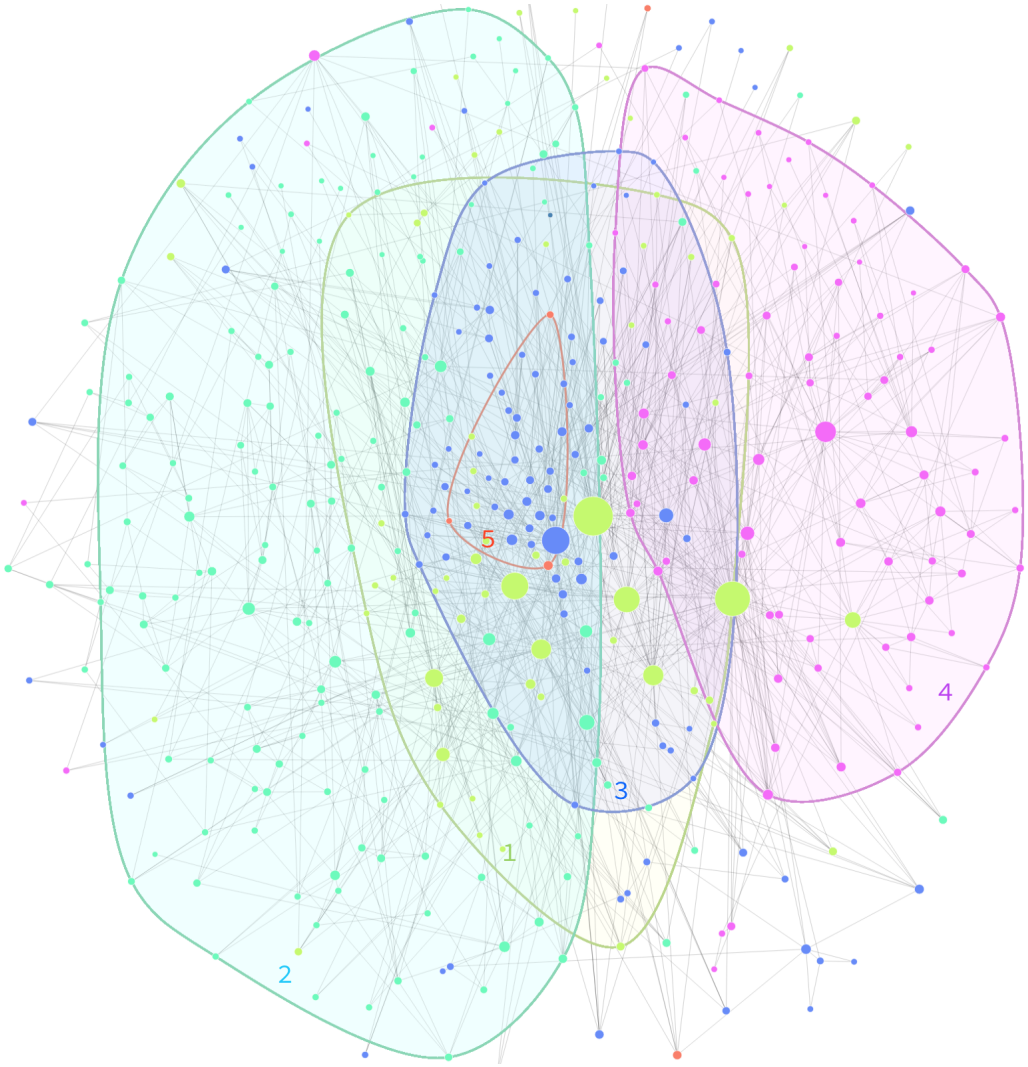


Figure 1.2:

Since the infomap provides a finer level of clustering which has originated from this, it is important to evaluate this. Using a graph-hull approach, as in Figure 1.2, becomes cluttered and unusable. Instead a bubble plot may be used. Although this sacrafices the ability to view links, it allows for the complete overview of the hierarchical structure. In Figure 1.3 shows the nested structure of each clustered group. In an electronic mail correspondance with **?** the origin of the naming convention of reduced species was explained. Using this, individual nodes are categoriesed by their prefixes. This allows the further categorisation of the chemistry withich has been grouped into a category.

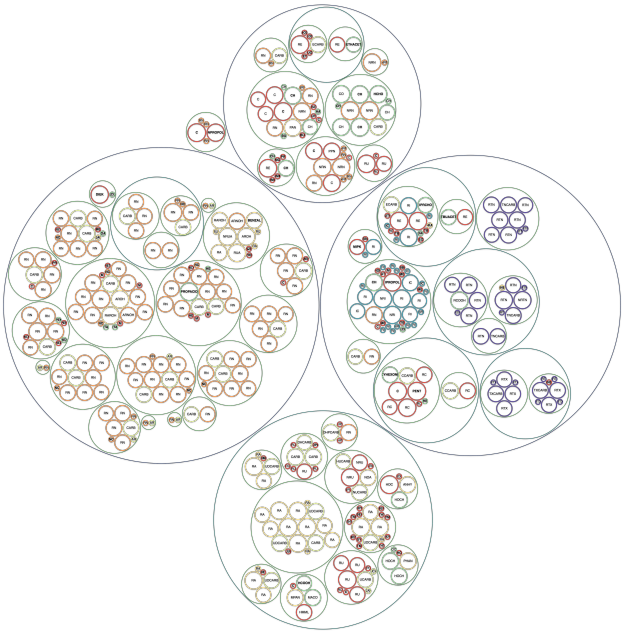


Figure 1.3: node sizes are representative of the $\log_{10}$ number of walkers that have become trapped by the flow algorithm at a location

1.4.4.1 Species type and clustering

Although a nested bubble chart is an intituive way to represent groups within a graph, a tree approach is more suited to revealing the hierarchical structure of the network. Figure 1.5 shows the same simulation represented in the form of a tree. Here branches are numerically labled on each level - this allows us to navigate the structure using a sequence of numbers, e.g. to get to 1.5.C_4H_6 we take the first branch from the centre, followed by the fith branch after that.

This split structure allow a general overview of the mechanism structure, as well as the reasoning/process of the clustering algorith. As was seen in Figure 1.2 the first level split shows branches 1,2 and 5 having origins from linear (n-) alkanes. This can be seen through both the emiited species (bold) and the *RN* prefix of the species. Here the linear alkanes can react with OH to extract a hydrogen and then from an RO_2, or produce a cabonyl *CARBxx*, which can then go on to produce the $RNxxO_2$ peroxy radical.

The with the exception of benzine in 2.14, branches 3 and 4 contain the aromoatic species in the network. Branches 4.{2,5,9,11} all consist of $RAxxO_2$ species, which are the procuct of the addition of OH to toluene/benzine ringed species. 4.{1,7,8} and 1.5 contain peroxy radicals formed from the degregation of conjugated dienes (two alkene groups separated by a single bond, where some sharing of electrons may occur) $RUxxO_2$. For the CRI v2.2 mechanism these are only isoprene and 1,3-butadiene. Such peroxy radicals often go on to form unsaturated carbonyls, as denoted by *UCARBxx*.

Branch 3 contains the monoterpenes. This can be seen in 3.{2,5} ($\alpha-$pinene) and 3.6 ($\beta-$pinenen).

Here peroxy radicals formed from the reaction with the e**n**docyclinc[5] and e**x**docyclinc[6] double bonds of $\alpha-$ and $\beta-$ pinene are denoted with the prefix *RTN* and *RTX*.

The $RIxxO_2$ prefix was originally used for the peroxy radicals iso ('i-') alkanes and their cabonyl products - branches 3.{1,4}, however they tend to mainly be used for smaller branched precursors which produce acetone (CH_3COCH_3) as a majour product in their oxidation chain (branch 3.1). As acetone is a particularly unreactive carbonyl, the fact that it is water soluble means that they may be washed out of the atmosphere by precipitation, [Andersson-Sköld et al., 1992]. This may has been seen to interrupt the ozone formation process under regional-scale photochemical smog conditions in north-western Europe [- from M.Jenkins PAPER? do you know what this is].

Finally, since the CRI index is representative of the oxidation potential it is common to see speacies containing the CRI value within a cluster. Cluster typically contain a combination of carbonyl (R(=O)R',$CARBxx$), hydroperoxy (R−OOH,$RxxOOH$), peroxy ($ROO\cdot$, $RxxO_2$) and nitrate (R−ONO_2,NO_3) groups. For the lumped species, it can be common for an RO_2 species to react with NO or NO_3 to produce a carbonyl with a CRI index of two values lower. This can be attributed to the loss of an oxygen and the formation of a double bond? (what is the long reaction for the MCM, it seems les direct). Similarly a reaction with NO or HO_2 can produce a hydroperoxy or nitrate species, which in turn react with OH to produce the an equivalent carbonyl.

1.4.4.2 Inter and Intra links

Typically the quality of clustering can be assessed using the adjacency matrix of a graph. In sortting the axis by cluster groups, squares of high density inter connections should become apparent in an adjacency heatmap. Unfortunately for an atmospheric chemistry mechanism, the size and sparseness of a graph, makes this an infeisable method to visually acess the nodcluster node ratios. Instead we can adapt the nodes presented by the treemap in Figure 1.5 and replacing the hierarchical structure, with one representing the original graph.

- greatest group, dark

- inter group to green - these are central on graph plot. important smaller

[5] Inside the pinene ring.

[6] Outside the pinene ring.

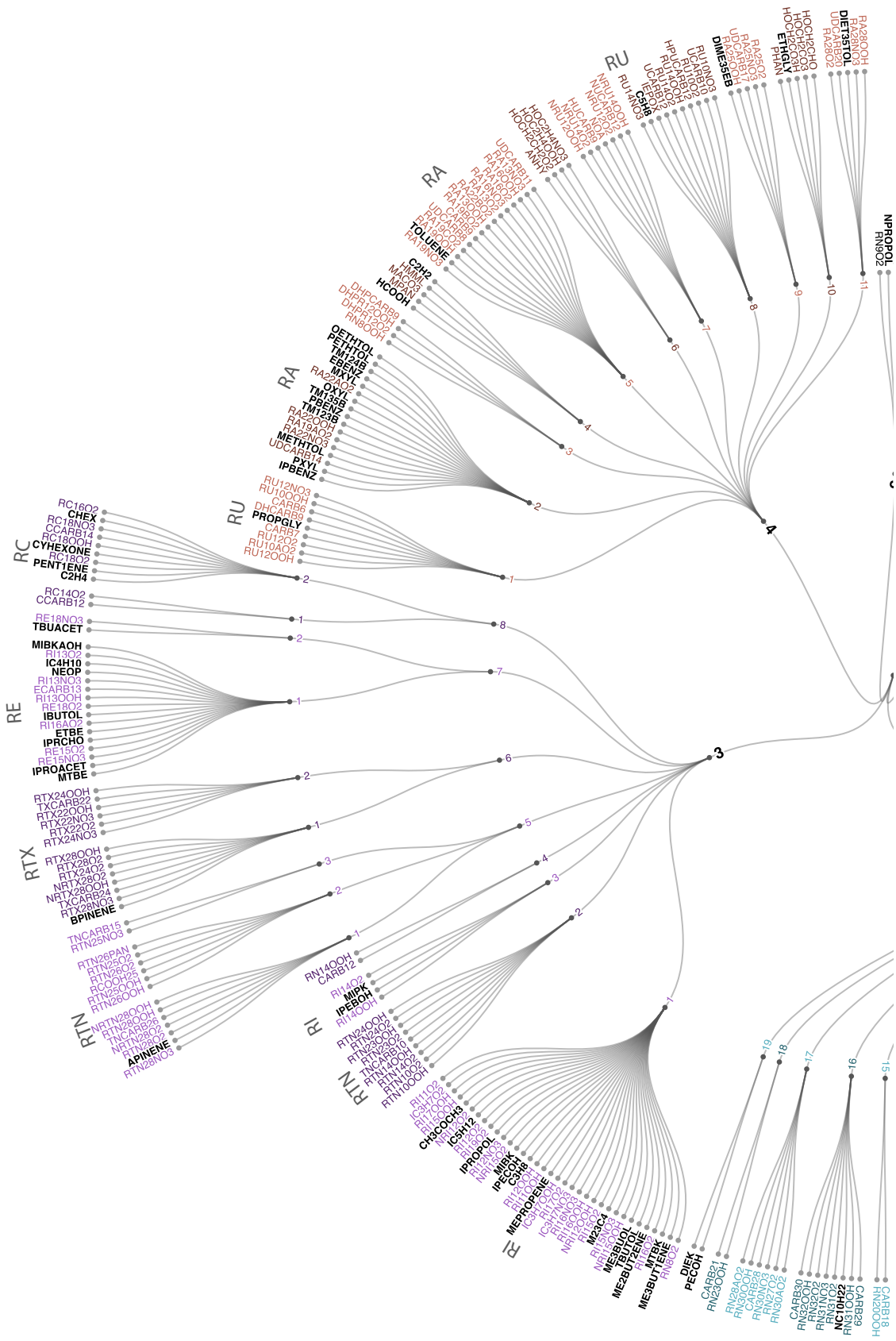
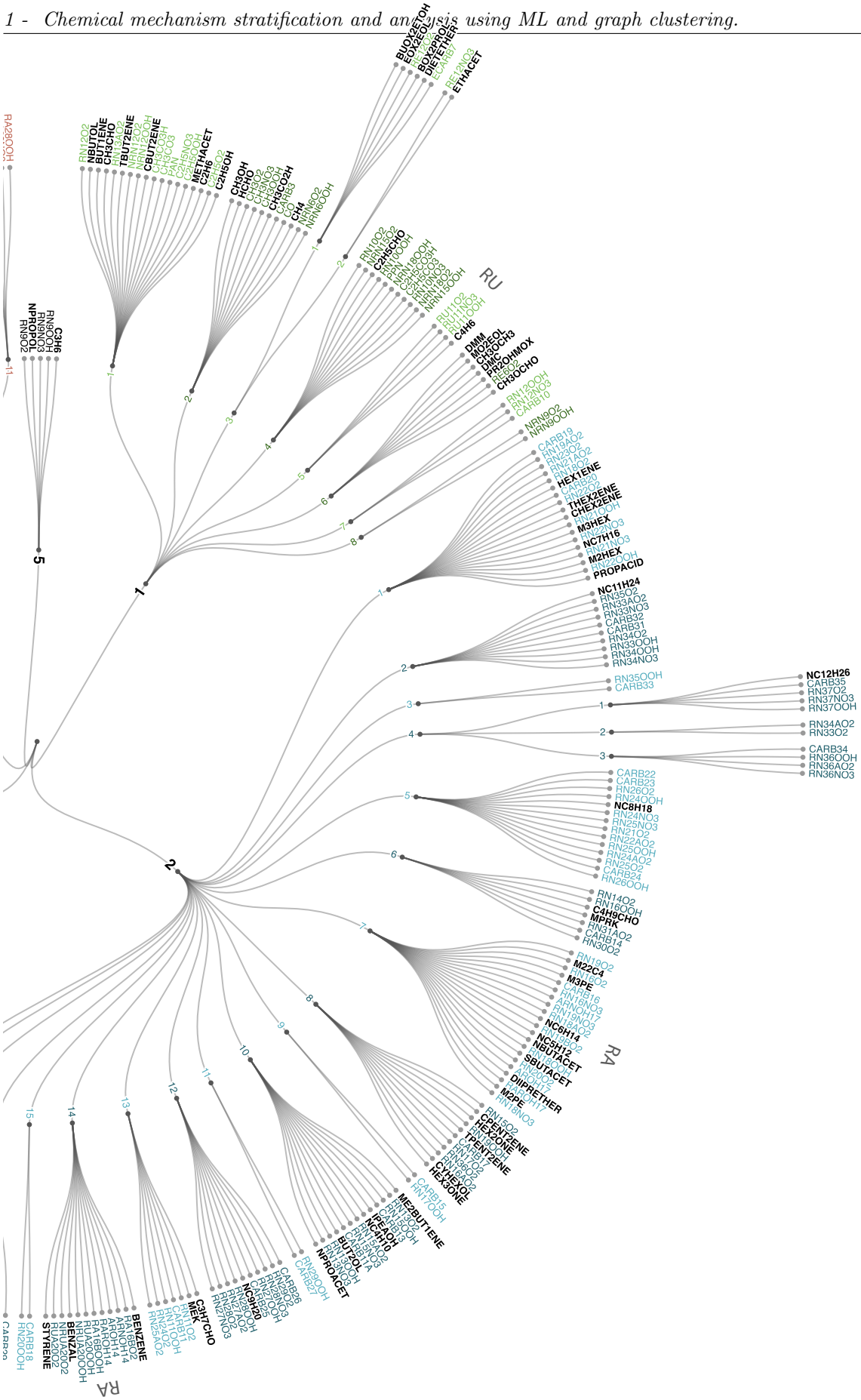


Figure 1.5: **A radial treemap showing the hierarchical clustering of the CRI mechanism.** The simulation results used are representative of the chemistry within London at Noon localtime and generated using DSMACC and the infomap algorithm.


RU
RA
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
C4H6
DMM
M02EOL
CH3OCH3
DMC
PR2OHMOX
RE6OO
CH3OCHO
NC7H16
M3HEX
M2HEX
HEX1ENE
THEX2ENE
CHEX2ENE
PROPACID
NC11H24
NC12H26
NC8H18
C4H9CHO
MPRK
M22C4
M3PE
NC6H14
NE5H12
NBUTACET
SBUTACET
DIIPRETHER
M2PE
C6ENT2ENE
HEX1ONE
TPENT2ENE
CYHEXOL
HEX3ONE
M2BUTTENE
IPEHEXOL
NC4H10
BUT2OL
NPROPACET
NC9H20
C3H7CHO
MEK
BENZENE
STYRENE
BENZAL
CH3OH
HCHO
CH3OOH
CH3NO3
CH3CO3H
CH3CO2H
CH4
NRN6O2
NRN6OOH
METHACET
C2H6
C2H5OOH
C2H5OH
C2H5CHO
C2H5NO3
C2H5CO3H
C2H5CO2H
NBUTOOL
BUTTEOL
CH3CHO
TBUT2ENE
CBUT2ENE
BUT2OL
BUT2PROL
BETETHER
ETHACET

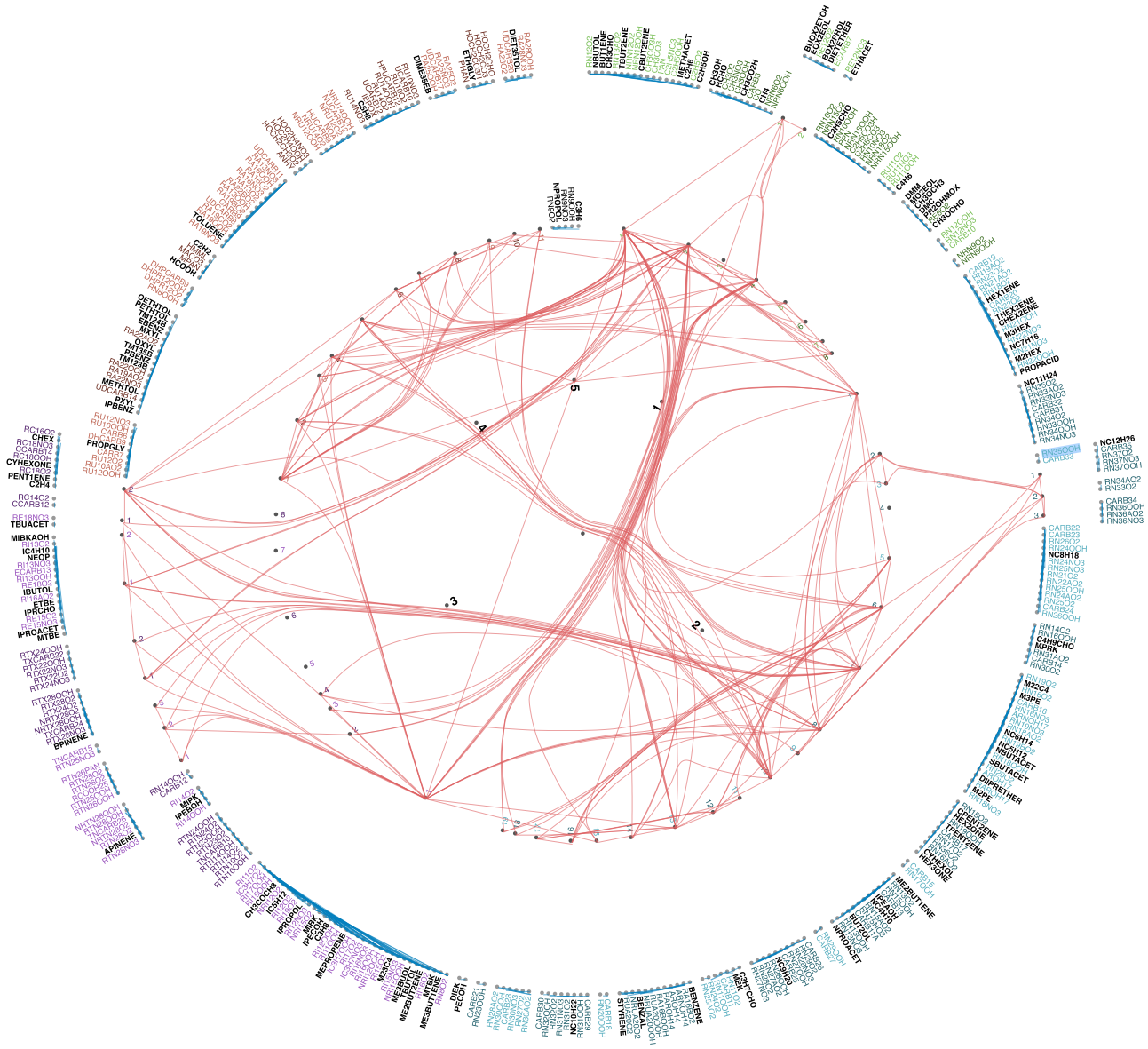


Figure 1.4: **Treeplot.**

1.4.4.3 Number of clusters

Sometimes it may be required to have slightly smaller or larger numbers of clusters. The infomap algorithm contains a *preffered number of modules* parameter which can either terminate the algorithm early, should the number be reached, or continue splitting if it has not. Since we are interesed in merging smaller numbers of nodes together, this can be seen as a useful parameter to have. However, in selecting a number too large, (e.g. 200 clusters, which should result in groups of 2-3 nodes), it is seen that much of the hierarchical information from the network is lost, Figure 1.6. It is for this reason that forcing the number of nodes without reason will not be attempted.

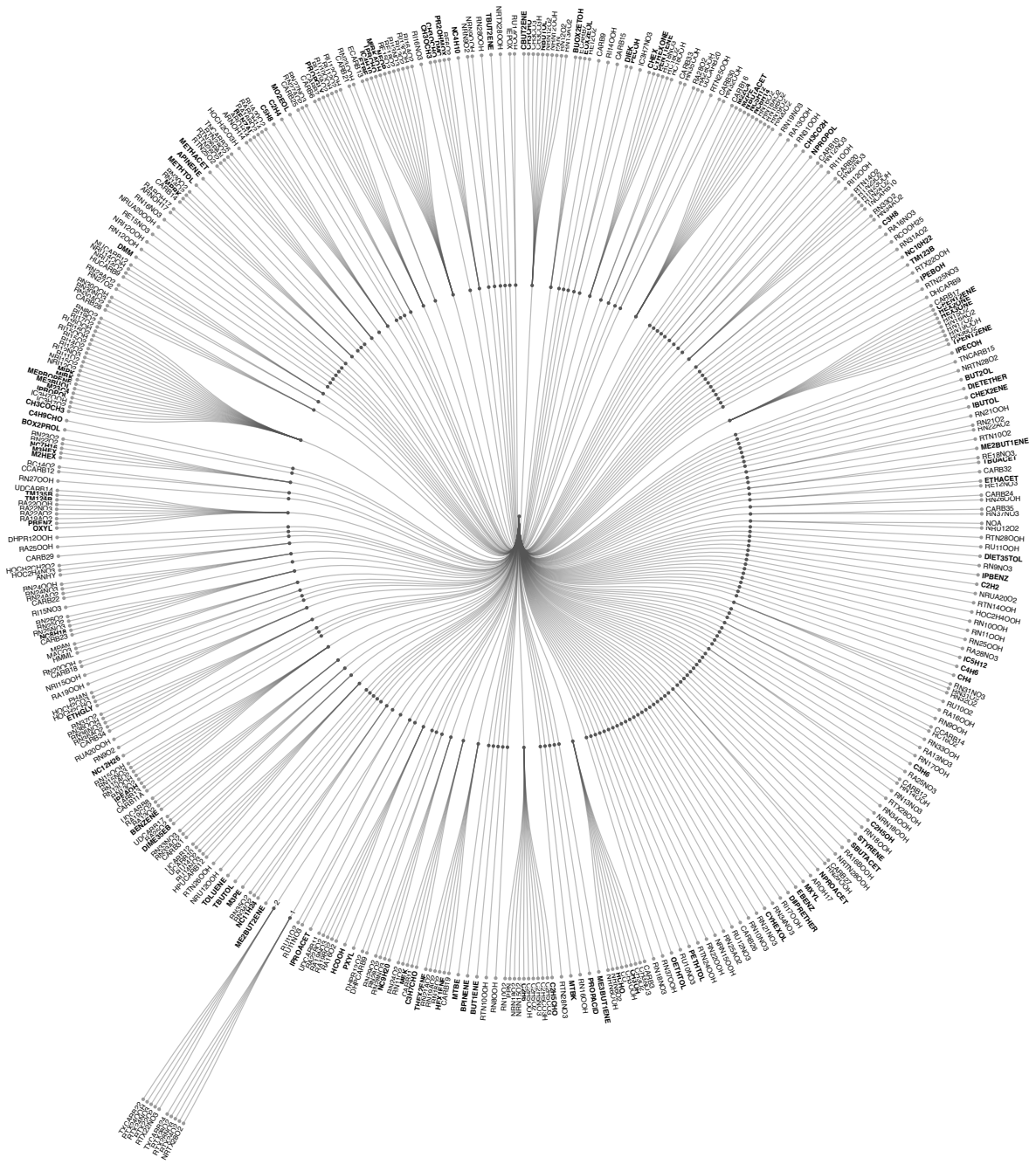


Figure 1.6: **A radial tree of the infomap algorithm with a forced number of groups.**

1.5 Reduction through Lifetime

Conventionally

1.5.0.1 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 1.5.0.1, 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

1.5.1 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.

1.5.1.1 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

$$v1 = [a, b, c, \dots n]$$
$$v2 = [i, j, k, \dots z] \tag{1.7}$$

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

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

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

1.5.1.2 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!).

1.5.2 Temporal Lifetime Vector Comparison

To compare a species diurnal profile with its absolute lifetime we can plot the cosine and eudlidian distance against eachother. In this subsection we compute the euclidean and cosine distances for all remaining reaction pairs (88410 pairs) for a single simulation. We start by looking that the species density profiles, Figure 1.7.

THIS Is the WRONG W|Ay ON KDE - matches distribution on plot!

Here we can see that most of the species within our mechanism fall under two peaks. The first is a peak consists of species which are mutally different in both diurnal profile and concentration change. This would be seen in species with inverse changes depending on the time of day. Although inorganic species have been ommited, an obvious candidate for this peak would be the similarity between NO and NO_2. The other large peak consists of species which have similar

The other contains species with a simiar dirunal profiles, that do not change at the same magnitude as eachother. In general the euclidean distances are smaller than the cosine distances, suggesting a greater concenration change between pairs of species with the same concentrtion profile shape.

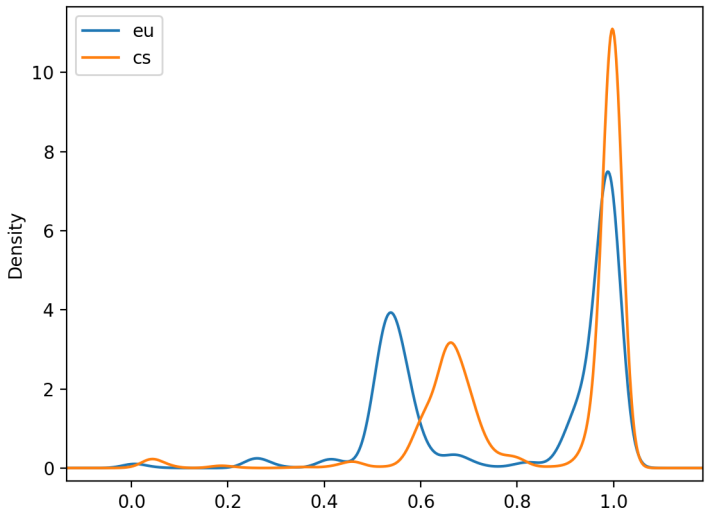
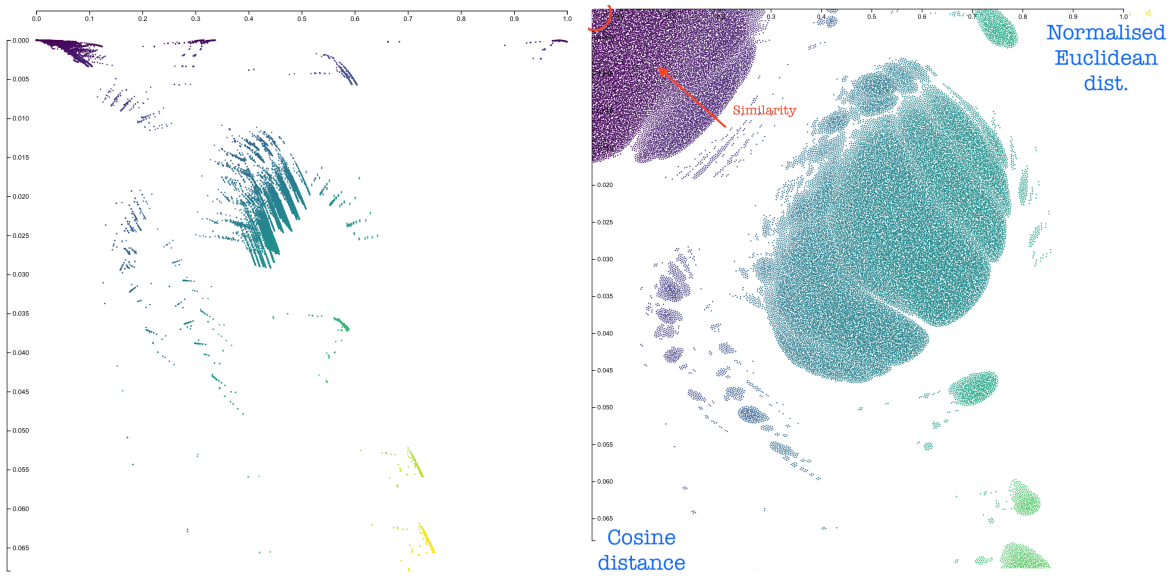


Figure 1.7: Gaussian Kernel Density Estimate plot showing the distributions present for the {0,1} scaled euclidean and cosine distances.

Next it is possible to compare both distance metrics on an $x - y$ plot - Figure 1.8a. As many species hace similar lifetimes, these are often situated within the same temporal space, which can make it hard to visually or interactively separate them. To overcome this, it is possible to convert the scatter plot into a force-simulation, Figure 1.8b. Here nodes repulse eachother, and are attracted to their original location. This expands the graph, and prevents overlapping nodes. In doing so it is possible to interactively query the paris of nodes which are represented by each point.



(a) Original (b) No-overlap

Figure 1.8: Showing the evolution from the original overlaid locations, Figure 1.8a to the slightly more accessible (interactively) Figure 1.8b

Using a Kernel density plot it is seen that both cosine and euclidean distances have a similar distribution of points for the chosen simulation. The agreement of both metrics suggests a similarlarity between both the lifetime values and their change over time for a simulation. The distances locate a set of species

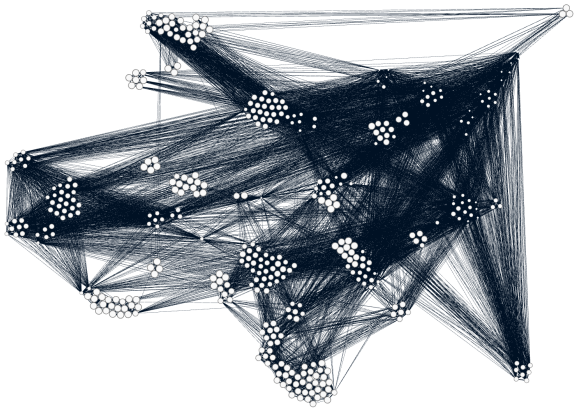
1.6 Results

In order to get a representation of the mechanism we run 300 randomly initiated scenarios. The experiemental setup is one such that it is possible to add more datapoints at a later date. From each simulation the no diagonal elements of the jacobian are used to construct a graph representative of the aggregated hourly means of the simulation output. Each of these graphs is then run through the infomap algorithm and a grouping/clustering produced. To select the best possible grouping, each infomap is run 100 times, where the result with the best fit (shortest codelength) is taken - this is an optional parameter on the algorithm.

1.6.1 The co-grouping network

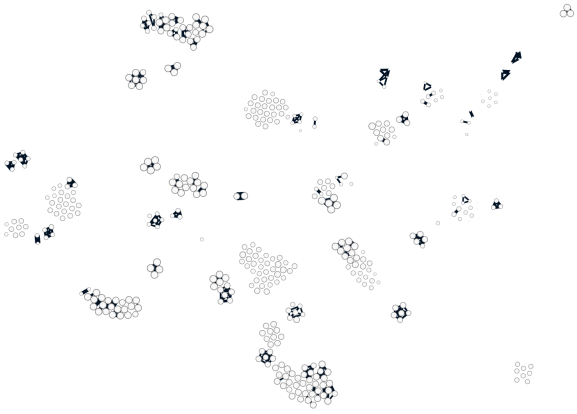
To aggregate the groupings produced by each algorithm an $n \times n$ matrix is created for each of the n species in the mechanism. This is treated as a graph relational matrix, whereupon if species A is in the same group as species B, then a link (or value +1) is added to the [A,B] (A $\longrightarrow$ B) and [B,A] (B $\longrightarrow$ A) column. Using this matrix format it is possible to then generate a graph showing the relationship between species that were clustered in the same group.

Using this matrix it is then possible to create a network, Figure 1.9a, which can progressively be filtered to represent only the nodes which are consistently paired together. Using prior knowledge that changes in chemistry follow a daytime and nightime regeme, we select only relationships that appear in over 45% of all the infomap runs.



(a) Full Graph

(b) >10% of graphs



(c) >20%of graphs

Figure 1.9: **Filetering the infomap clustering relationship matrix/graph** How the clutering relationship network changes as weak links (links between species which do not appear in many of the infomap groupings) are removed.

1.6.2 Comparing daytime and nightime groups

Having determined a set of species which are continuously clustered for over 45% of our simulation results, it is next important to see how these compare against those as part of a dirunal split. To this the output of the infomap clusters for midday and midnight are extracted. We then make use of an alluvial diagram. This is a cross between a parallel line plot and a sanky diagram and is often used to show changes in categorical data. This makes them particularly suitable for showing the changes of clusters within a temporal networks, as has been shown in [Rosvall and Bergstrom, 2010].

Figure 1.10 shows the

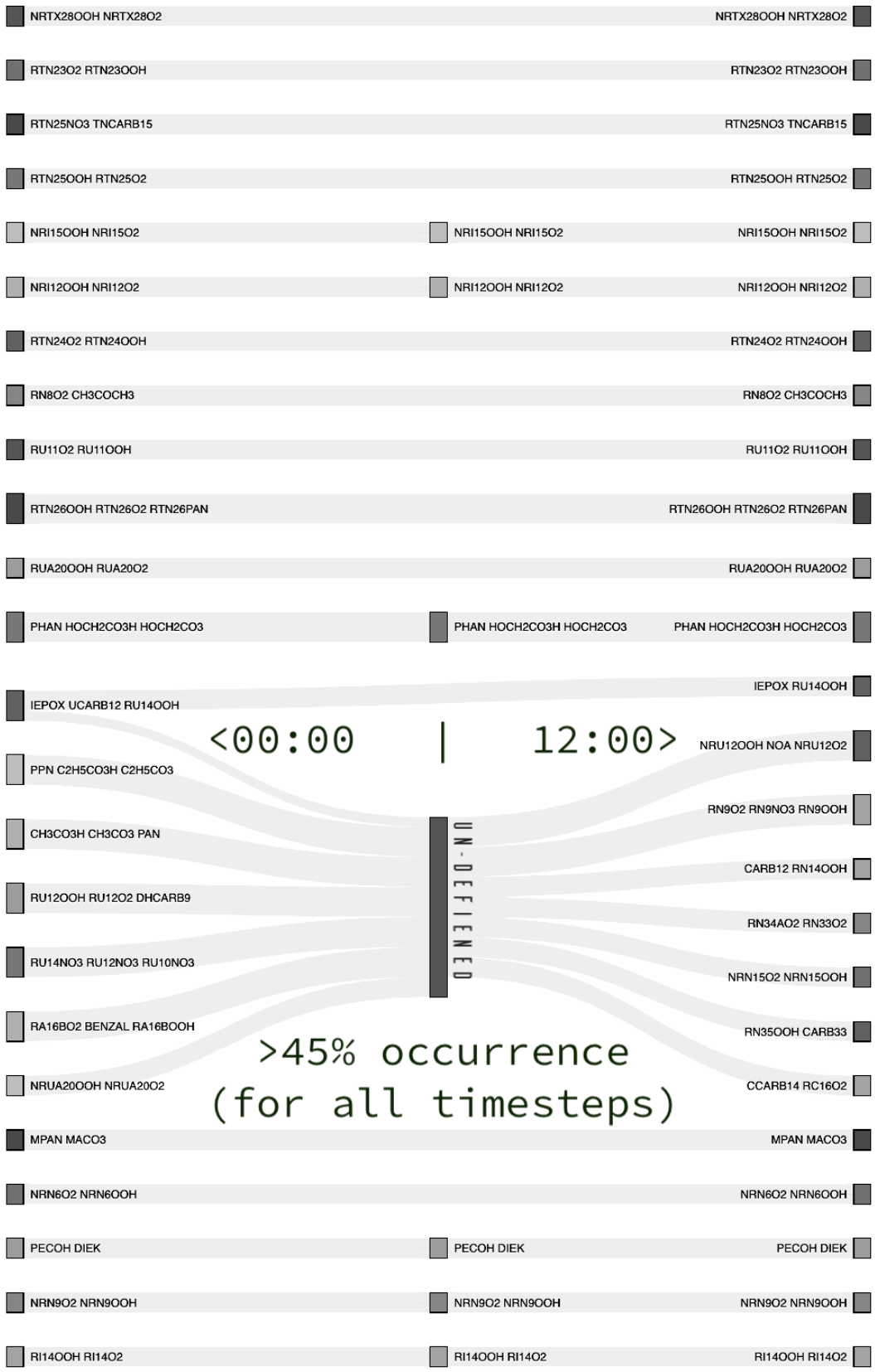


Figure 1.10: **An alluvial diagram showing the changes in clusters between noon and midnight.** On the left are all groups that apear in >45% of the midnight simulation results. On the right are groups which appear >45% of the midday results. In the middle exist the clusters extracted which appear in >45% of all runs. Here it is seen that there exist a series of species which may exist in daytime or nighttime chemistry, but do not persist between both.

1.6.3 Determining cluster suitabiltiy

Similarly the lifetimes of each species are extracted from the diagonal of the jacobian matrix. Using these the euclidian and cosine similarities are calculated and can be used to determine how well suited a species pair is for beign lumped together.

Bibliography

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

Andersson-Sköld, Y., Grennfelt, P., and Pleijel, K. (1992). Photochemical Ozone Creation Potentials: A Study Of Different Concepts. *Journal of the Air & Waste Management Association*, 42(9):1152–1158.

Aslak, U., Rosvall, M., and Lehmann, S. (2018). Constrained information flows in temporal networks reveal intermittent communities. *Phys. Rev. E*, 97:062312.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Cornes, P. (2008). Proposed Plans For Holme Pierrepont Whitewater Course. `https://hppconcern.wordpress.com/2008/08/04/proposed-plans-for-holme-pierrepont-whitewater-course/`. Accessed: 2020-2-15.

Everett, M. G. and Borgatti, S. P. (1994). Regular equivalence: General theory. *The Journal of Mathematical Sociology*, 19(1):29–52.

Fortunato, S. (2010). Community Detection In Graphs. *Physics reports*, 486(3):75–174.

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.

Lu, H., Halappanavar, M., and Kalyanaraman, A. (2015). Parallel Heuristics For Scalable Community Detection. *Parallel computing*, 47:19–37.

Mahajan, S. (2008). The Art Of Approximation In Science And Engineering. *MIT OpenCourseWare*.

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.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113.

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*.

Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76:036106.

Rosvall, M., Axelsson, D., and Bergstrom, C. T. (2009). The map equation. *The European Physical Journal Special Topics*, 178(1):13–23.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123.

Rosvall, M. and Bergstrom, C. T. (2010). Mapping Change In Large Networks. *PloS one*, 5(1):e8694.

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.

Zhou, H. (2003). Distance, dissimilarity index, and network community structure. *Phys. Rev. E*, 67:061901.