UvA-DARE (Digital Academic Repository)

Graph-theoretical approach to algorithmic construction of complex reaction networks

Orlova, Y.

Publication date
2020
Document Version
Other version
License
Other

[Link to publication](#)

Citation for published version (APA):
Orlova, Y. (2020). *Graph-theoretical approach to algorithmic construction of complex reaction networks*.

2

Automated reaction generation for polymer networks

Most theoretical studies of polymer kinetics are performed by manually reducing the chemical system to a few simple reaction mechanisms. Not being constrained by such reducibility, this work considers the polymerization as a product of a complex network of reactions that need not to be known in advance. Combining various ideas from graph theory, combinatorics and random graphs, we introduce a new modeling approach to complex polymerization that automatically constructs a reaction network, solves the kinetic model, and retrieves such topological properties of the final polymer network like, for instance, distribution of molecular weight. In this way, the new approach acts as an intermediate layer that propagates the knowledge of the basic chemistry in order to capture and understand the complexity of the real world polymerizing systems.

2.1. Introduction

Many real-life chemical systems consist of a large number of molecular species perpetually reacting with each other. Writing the formal reaction mechanism by hand, even for small molecules, is a labor-intensive task. One molecule can undergo numerous transformations until it reaches its final state. In some special cases, as for instance in polymers consisting of interconnected repeat units, even the notions of a species or the final state are not well defined, as they are rather expressed as probability distributions. This leads to extraordinarily large networks of interdependent reactions, counting thousands of distinct chemical species and much larger numbers of reactions. This proved to be an interesting topic for a wide range of application areas. To name a few, a theory explaining the origin of life as arising from simple organic compounds[1] and a number of plausible prebiotic scenarios[2] have been developed with the help of reaction networks. Analysis of metabolic networks[3],[4],[5],[6] allows to study gene regulation, information transfer, cell-fate decision and many other important processes associated to the cell cycle. Furthermore, reaction networks are applied to the field of drug discovery[7] and the risk assesment of hazardous byproducts[8]. Various useful properties of chemical systems can be extracted solely from the underlying reaction network, as for instance, absolute concentration robustness[9]. The present chapter offers a novel methodology to design reaction networks in the context of complex polymerization processes.

In areas other than polymerization kinetics the complexity of real life reaction systems has motivated the development of many software packages to automatize reaction mechanisms. For example, MAMOX[10] sofware generates detailed kinetic models for pyrolysis, oxidation or combustion mechanisms of large hydrocarbons. EXGAS[11] is developed to model the oxidation of linear and branched alkanes up to C_{16}, linear alkenes from C_3 to C_7 and cycloalkanes. NetGen[12] reconstructs chemical reaction networks for gas phase pyrolysis, biochemical reactions and nanoparticle synthesis. This package also allows to perform on-the-fly quantum chemistry calculations to estimate the kinetic parameters of the system. Reaction networks of free radical hydrocarbon chemistry are well reconstructed by RMG[13]. The algorithm generates all possible molecular species and reactions and then estimates whether the reactions are thermodinamically feasible and should be added to the final mechanism. Genesys[14] is a rule-based reaction generating software which also allows the estimation of the kinetic parameters. In this package, the reaction rules are defined by the user, which allows a large flexibility for its use in a broad range of chemical systems. RING[15] reconstructs reaction networks from a mechanism written in English-like reaction language. The software also offers many graph theory tools for further analysis of the resulting reaction network.
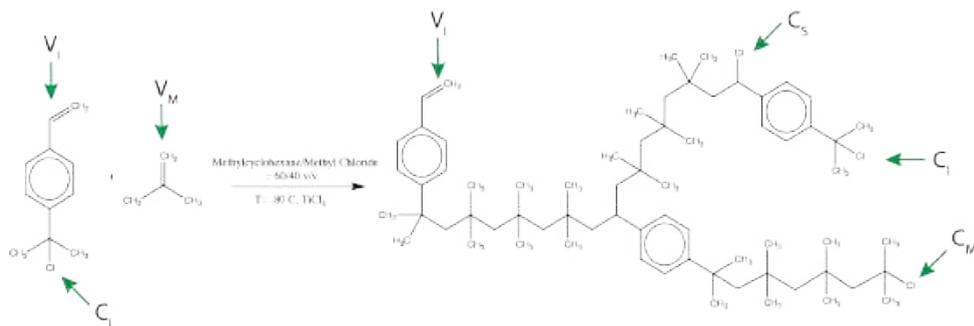
Figure 2.1: An example of a polymer network formed due to copolymerization of IM and IB. V_I, V_M, C_I, C_M and C_S are functional groups, which are responsible for the polymer network formation.

In the polymerization processes, monomers gradually connect to each other forming higher order arrangements: linear chains or networks. Progress was made in modeling of polymer microstructure using advanced Monte Carlo techniques[16–19]. However, from the perspective of currently available network-generation packages, polymer networks are treated as large molecules and thus the repeatability of structural units is being ignored. As the size of a molecule grows, it becomes rather complicated to capture all possible intermediate products and the reaction pathways. Graph theory tools which are the foundation of most reaction generating mechanisms, become computationally prohibitive due to NP-hardness: there are no polynomial time algorithms which can solve some of the graph theory problems[20]. Moreover, the reaction network of the polymerization process has infinite size, which is the main limiting factor.

Due to such a prohibitive computational complexity, currently available case studies of polymerization regarded only well-defined reaction mechanisms. That is to say, the mechanisms in which both the repeated units and the complete set of reactions are known in advance. In these cases, the numerical computations are of polynomial complexity and are necessary only to perform time integration of the kinetic model in deterministic or stochastic fashions. Efficient tools have been developed describing macromolecular properties in terms of numbers of monomer units, branches, functional groups, etc.[21–25]. A class of polymerization systems exists, however, with more than usually complex chemistry, for which we nevertheless desire to predict macromolecular properties. An example is the drying of linseed oil[26,27], where a complex radical-based photo-oxidation process leads to the formation of a dense polymer network. Properly dealing with such a system would require combining the best of both worlds of polymer modeling and automated reaction network generation. The present chapter introduces

a new methodology designed for complex polymer systems, and constitutes a synergy between these two, rarely intersecting, modeling worlds.

The overview of automated reaction network generation shows that many concepts and algorithms have already been developed. These have proved to be valuable in the development of our new methodology. Similarly to the above-described reaction generating packages, we utilize a reaction-rule approach to generate the reaction network. Knowing the reaction families and initial species, our method reconstructs all the intermediate and product molecular structures, while keeping track of the reactions that occur between them. The novelty of our method is - unlike all the previous developments - that we keep the model on the level of repeated units and therefore limit the size of molecular species involved. This means that the resulting polymer network is defined only by its fragments. In the terms of our methodology, we will refer to these fragments as "species". Although in this way the explicit macromolecular topology does not directly become available, the essential features of the topology may be retrieved. For instance, the size distribution of connected components can be reconstructed by using tools from random graph theory[28–30]. Other topological properties of the polymer networks that can be obtained from the degree distribution are discussed in the work by Kryven[31].

We illustrate the methodology for a reaction generation mechanism on the example of self-condensing vinyl polymerization with 4-(1-hydroxyl-1-methylethyl)styrene-co-styrene as a macro-initiator and $TiCl_4$ as a Lewis acid in a batch reactor, which produces finite macromolecules. More details about chemistry can be found in[32]. This reaction mechanism is well studied in the work by Zhao *et al.*[33] by means of advanced Monte Carlo modeling. The reaction equations listed in Table 2.1 are expressed in terms of the reactive groups V_i and C_j, which denote the functional groups as shown in Figure 2.1. This figure shows a fragment of the polymer network, which can be formed due to copolymerization of inimer (IM) and isobutylene (IB). Since the IM/IB copolymerization has a relatively simple kinetics, it serves as a good context to introduce our new methodology. The rate parameters are known from the literature, so we do not deal with parameter estimation in this methodology. At the same time, the methodology itself is not bound by the scope of this chemical system and constitutes a general framework for complex polymerization.

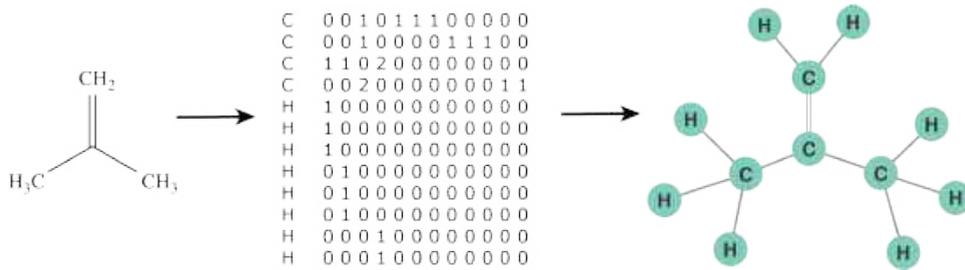Table 2.1: Reaction steps for copolymerization of IM with IB, rate coefficients are in $Lmol^{-1}s^{-1}$

Figure 2.2: Molecular graph representation of an isobutylene molecule.

$(^{33})$

#	Reaction	Rate coefficient
1	$C_I + V_I \rightarrow C_S$	3.32×10^{-2}
2	$C_I + V_M \rightarrow C_M$	4.46×10^{-4}
3	$C_M + V_I \rightarrow C_S$	5.19×10^{-1}
4	$C_M + V_M \rightarrow C_M$	2.27×10^{0}
5	$C_S + V_I \rightarrow C_S$	6.45×10^{-3}
6	$C_S + V_M \rightarrow C_M$	4.11×10^{-5}

This chapter is organized as follows. After this Introduction, in Section 2 we introduce the definitions and tools necessary for the description of the reaction generating mechanism. In Section 3 we describe the general idea of the algorithm, from which the complete reaction network is obtained. Main steps of the methodology are illustrated in Section 4 by its application to copolymerization of inimer and isobutylene in a batch reactor.

2.2. Tools to describe the reaction system

In this section we will introduce various concepts, definitions and tools required to describe the reaction generating mechanism for the case of polymerization.

2.2.1. Graph theoretical background

A graph $G = (V, E)$ is an abstract data structure, which consists of a set of n nodes (or vertices) $V = \{v_1 \ldots v_n\}$ and a set of edges $E \subseteq V \times V$, which determines the connections between nodes. The size of a graph G is n - the number of nodes in a graph. A graph G is undirected if $(v_i, v_j) \in E$ implies $(v_j, v_i) \in E$, otherwise a graph is directed. A graph is labeled if each node has a corresponding label $l(v_i)$. If a graph is labeled, then it is defined by a triplet $G = (l, V, E)$. The edges can be labeled as well.

A graph G is also described by its adjacency matrix $A \in \mathbb{R}^{n \times n}$. In the case of an undirected graph, a_{ij} has a value of 1 if there exists an edge between nodes v_i and v_j, otherwise a_{ij} has a value of 0. In the case of a directed graph, a_{ij} has a value of 1 if an edge points from node v_i to node v_j, otherwise a_{ij} has a value of 0.

Nodes v_i and v_j are adjacent, if there exists an edge between them: $(v_i, v_j) \in E$. The neighborhood $N(v)$ of a node v is the set of all nodes which are adjacent to v: $N(v) = \{v_i | (v, v_i) \in E\}$. The degree $deg(v)$ of a node v is the number of its neighbors, or nodes v is adjacent to. If a graph G is directed, the degree of a node v can also be classified according to the type of edges connected to v. $deg_{in}(v)$ is the incoming degree: number of neighbors which are connected by incoming edges to v and $deg_{out}(v)$ is the outgoing degree: number of neighbors which are connected by outgoing edges to v.

The first order neighbors $N_I(v)$ of a node v coincide with the definition of node neighborhood. In other words, the first order neighbors of v are nodes, which lie at a distance of one edge from v. The second order neighbors $N_{II}(v) = N(N_I(v))$ of a node v are "neighbors of neighbors": nodes, which lie at a distance of two edges from v. A graph $G = (V, E)$ is bipartite, when a set of nodes V can be split in two disjoint sets U and W, such that for all edges $(v_i, v_j) \in E : v_i \in U, v_j \in W$. For a bipartite graph, if $v \in U$, then $N_I(v) \subseteq W$ and $N_{II}(v) \subseteq U$.

Two graphs $G = (l, V, E)$ and $G' = (l', V', E')$ are called isomorphic, if there exists a bijective mapping $f : V \rightarrow V'$ such that $(v_i, v_j) \in E$ if and only if $(f(v_i), f(v_j)) \in E'$ and $l(v_i) = l'(f(v_i))$ for each node $v_i \in V$.

A major part of the objects and their relationships used in this methodology can be described in terms of graphs. In this chapter, we distinguish three types of graphs: 1) molecular graphs, 2) reaction network graphs and 3) graphs representing the polymer network. Obviously, the nodes and edges in these three different graphs refer to different entities, as will be explained when we discuss the role of the graphs. More information on graph theory can be found in [34].

2.2.2. Molecular graphs

The chemical species are represented by molecular graphs. This representation does not contain the positions of atoms in 3D space within a molecule.

A molecular graph is composed of its adjacency matrix A and list of labels l. The list of labels consists of the names of the chemical elements which correspond to atoms in a molecule. For a molecule of n atoms, the adjacency matrix A of size $n \times n$ contains connectivity and the bond order information. Entry of the adjacency matrix $a_{ij} = 0$

implies that atoms i and j do not share a bond. Single and double bonds between atoms i and j are indicated by $a_{ij} = 1$ and $a_{ij} = 2$ respectively in an adjacency matrix. A radical atom i is indicated by $a_{ii} = 1$ in the diagonal of the matrix A. Figure 2.2 illustrates the molecular graph of isobutylene.

2.2.3. Monomer approach and crosslinks

In a polymerization process, monomers connect to each other as a result of various reactions to form bigger structures: connected components. The molecular graph of the connected component will rapidly become excessively large during these reactions. Hence, in order to overcome this problem we introduce the "monomer approach". In this approach, the whole polymer network is parametrised by its degree distribution: the probability to find a monomer of specific type and number of connecting bonds to other monomer units at a given point in time. The size-distribution of connected components can then be inferred by applying random graph theory. Note that we will henceforth use the term "crosslinks" for these connecting bonds. Obviously, these bonds between different monomers are chemically indistinguishable from bonds inside monomers. Also, the reader should not confuse these "crosslinks" to bonds formed by a specific class of branching reactions, in particular in crosslinking polymerization.

We thus do not store information on the whole polymer components, but only the monomers with different configurations of adjacent crosslinks which are present in the structure of these components. Figure 2.3 shows the coarse-grained representation of the polymer network previously introduced in Figure 2.1 for IM/IB copolymerization, showing monomers with one, two or three links to other units, so having degree u = 1, 2 or 3. In the box with crosslinked species, the set of unique monomers with adjacent crosslinks is demonstrated.

The types of crosslinks depend on the nature of the crosslinking reactions. Most of the crosslinking reactions can be divided into two pools: directional and non-directional (symmetric). The crosslink is considered to be symmetric if it is formed due to a radical-radical crosslinking reaction. Thus, if two monomers sharing a neutral crosslink are found, one can conclude that each of them carried a radical before the reaction happened. On the other hand, the crosslink is directional when, for example, a radical attacks a vinyl group. In this case, it is important to know which monomer carried a radical and which monomer had a vinyl group before the crosslink formation. Such a crosslink is represented as a directional edge, which points out of the former radical species to the species with the former vinyl group. Figure 2.3 contains directional crosslinks, represented as edges (crosslinks) between nodes (monomers) for the case of
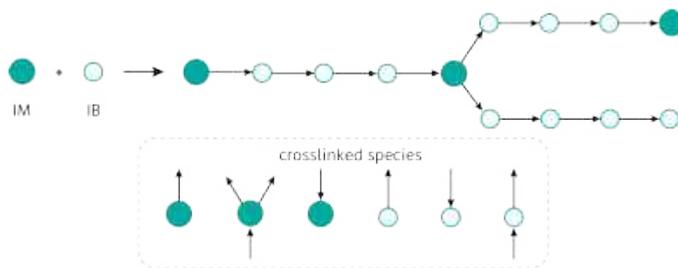
Figure 2.3: Example of a coarse-grained representation of the polymer network, shown in Figure 2.1, where each node is a monomer. Monomers are connected by directed crosslinks. Monomers with different configuration of incoming and outgoing crosslinks are illustrated in the box below the network. The network can also be described by counting how many monomers with different configuration of adjacent crosslinks are present in the system.

IM/IB copolymerization.

In the molecular graph notation, crosslinks are placed on the diagonal of the adjacency matrix. A different alphabet is used to indicate crosslinks. Since labels 1 and 2 are reserved to indicate chemical bonds and radicals, $3,4,5\ldots$ is used to represent different types of crosslinks. For example, 3 represents an atom which was a radical and then formed a crosslink with a vinyl group of another molecule. The former vinyl group is indicated by 4. Therefore, having two species with indicators 3 and 4 on the diagonals of the respective adjacency matrices, implies that a dimer is formed by a radical attaching to a vinyl group.

2.2.4. Reaction rules and patterns

Comparing the molecular graphs of a species before and after a reaction, one can notice that only a small part of the whole graph is changed, and the rest of the molecule stays invariant. Based on this fact, we introduce reaction rules on subgraphs rather than on the whole molecular structure. A pattern is a fragment of a molecule, which undergoes changes due to a reaction. Patterns are usually the functional groups as they are mainly responsible for the reactivity of the molecule. In terms of the molecular graph representation, patterns are subgraphs of the molecular graphs. This concept has been also used in other reaction generating packages[13],[14],[12].

A pattern has to include all atoms that undergo changes due to a reaction and some neighborhood of them. For example, one functional group can undergo the same transformation, but the reaction happens faster or slower depending on the neighboring atoms. An example of patterns on molecules is shown in Figure 2.4. One molecule can have several patterns.

A reaction rule is a transformation applied to a pattern of a reactant to produce a pattern of a product. Patterns of reactants are usually found in more than one molecule. Thus, one reaction rule can cover a whole family of reactions. The reaction rules are introduced manually in this methodology. However, some reactions might be found in different chemical processes and can be used for more than one system.

Many chemical systems involve first and second order reactions. A first order reaction is determined by one reactant which is transformed into one or several products. Second order reactions require two reactants and produce one or several molecules as products. An example of a second order reaction is shown on Figure 2.4. The pattern of Reactant 1 loses its chlorine atom and gets an outgoing crosslink. The vinyl group in the pattern of Reactant 2 gets a chlorine atom and an incoming crosslink. Following the monomer approach, the product is still described as separate species.

2.2.5. Pattern matching

The presence of patterns in a molecular graph determines its availability for the corresponding reaction. Thus, it is important to recognize patterns in a molecular graph. In the context of this section, 'recognize a pattern' means to decide whether a pattern is the subgraph of a molecular graph. In other words, one needs to find a correspondence between nodes in the adjacency matrices of a pattern and a molecular graph. This is the subgraph isomorphism problem. A subgraph is referred to as a pattern and a reference graph is referred to as a molecule. Subgraph isomorphism is an NP-complete problem: the number of operations required to solve the problem grows exponentially with the size of the graph. However, it is possible to reduce the search space and eliminate a considerable number of node comparisons at an early stage by accounting for limitations imposed on the molecular graph by the chemical context.

Different algorithms are available to deal with subgraph isomorphism, an overview can be found in[35]. Ullmann's algorithm[36] is a robust and frequently applied graph isomorphism detection method. It is based on brute-force enumeration with refinement steps, which serves to cut out most of the non-feasible search pathways at an early stage. Ullmann algorithm aims to find all isomorphisms of a subgraph $G' = (l', V', E')$ and a reference graph $G = (l, V, E)$. The algorithm works with an $n \times m$ matrix M of boolean values, where n is the size of a subgraph and m is the size of a reference graph. An entry m_{ij} of matrix M is equal to "1" if a node $v'_i \in V'$ can be matched to a node $v_j \in V$ and "0" otherwise. The goal is to find matching matrices, where each column has no more than one entry of "1" and each row has exactly one entry of "1". The failure criterion for one pathway is when at least one row contains all zero entries.
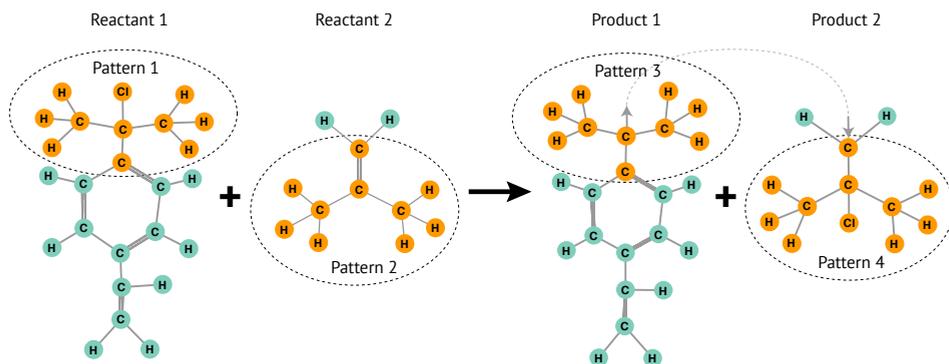
2

Figure 2.4: An example of the reaction rule for the initiation step of the V_M group an IB monomer by the C_I group on an IM monomer. Patterns, the fragments of a molecule responsible for the reaction, are colored in orange. The reaction also creates a crosslink, which is represented by incoming and outgoing arrows in the Products' patterns: an outgoing bond of type 3 for IM and incoming bond of type 2 for IB.

This means that at least one atom from the subgraph exists that cannot be matched to any node of the reference graph. At failure the process stops and backtracks to explore other possibilities for matching.

Ullmann suggested the following conditions under which nodes $v_i' \in V'$ and $v_j \in V$ can be matched:

- **Label and degree condition:** nodes can be matched if : $l'(v_i') = l(v_j)$ and $deg(v_i') \leq deg(v_j)$.

- **One to one mapping condition**: nodes, which are already mapped do no take part in matching till the end of the current search pathway.

- **Neighborhood condition**: a node from a subgraph can be matched to a node from a reference graph only if the neighbors of the subgraph node match to the neighbors of the reference graph node.

A detailed explanation of Ullmann's algorithm can be found in[36]. This version of the Ullmann algorithm is suitable for the general case of labeled graphs. In the case of molecular graphs, bonds, crosslinks and radicals provide this additional information for a matching step.[37] describe how Ullmann's approach can be adapted to chemical structures. Taking into account the matching criteria from[37], certain further "chemical" conditions can be checked on each matching step:

- **Radical condition**: radicals should be matched to radicals.

- **Bonds with neighbors**: the set of neighbors' labels of node $v'_i \in V'$ has to be a subset of the neighbors' labels of node $v_j \in V$ and these neighbors should have the same type of bonds with nodes to be matched.

- **Crosslink condition**: if node $v'_i \in V'$ has a crosslink, it can only be matched to $v_j \in V$, if v_j has the same type of crosslink as v'_i. However, if v'_i has a crosslink, it can still be matched to a subgraph node without a crosslink.

These conditions are helpful to check during the initialization of matrix M, as they can early eliminate some non-feasible search pathways at an early stage.

2.3. Algorithm development

2.3.1. Automated reaction mechanism

In the previous section we have discussed the tools required for the description of the reaction system. Here, we explain the algorithm of our new methodology that employs these tools to automatically generate the complete reaction mechanism: all the species and reactions involved. The algorithm starts with the input species and a small set of reaction rules including all the patterns. Thus, it develops the structures of all possible intermediate and final species, which can be found in the system after being exposed to the reaction rules.

Firstly, the initial species have to be converted into molecular graphs: adjacency matrix plus a list of labels (see Figure 2.2), and stored in the list of species. This list is going to be expanded by products from the reaction rules in an iterative manner.

The next step requires to go through the reaction rules to find candidate species for each reaction. First order reactions require only one candidate for a reactant, while two candidates have to be found for each reactant in a second order reaction. A reaction happens only if all of its reactant patterns have a corresponding candidate in the list of species. In the very beginning of the reaction process, only few reactions can happen because the list of species does not have many molecules. If for a given reaction rule candidate reactants are found among the species in the list, the transformation from reactant pattern to product pattern is applied to produce potentially new species.

During the transformation of a molecule from reactant to product only the part of the graph which corresponds to reactant pattern undergoes changes - as prescribed by the reaction rule - the rest of the graph remains unchanged. The product's graph is then obtained by removing atoms which correspond to the reactant pattern and replacing them by atoms from the product pattern, producing a new adjacency matrix that corresponds to a product species and may not be present in the list of species yet.
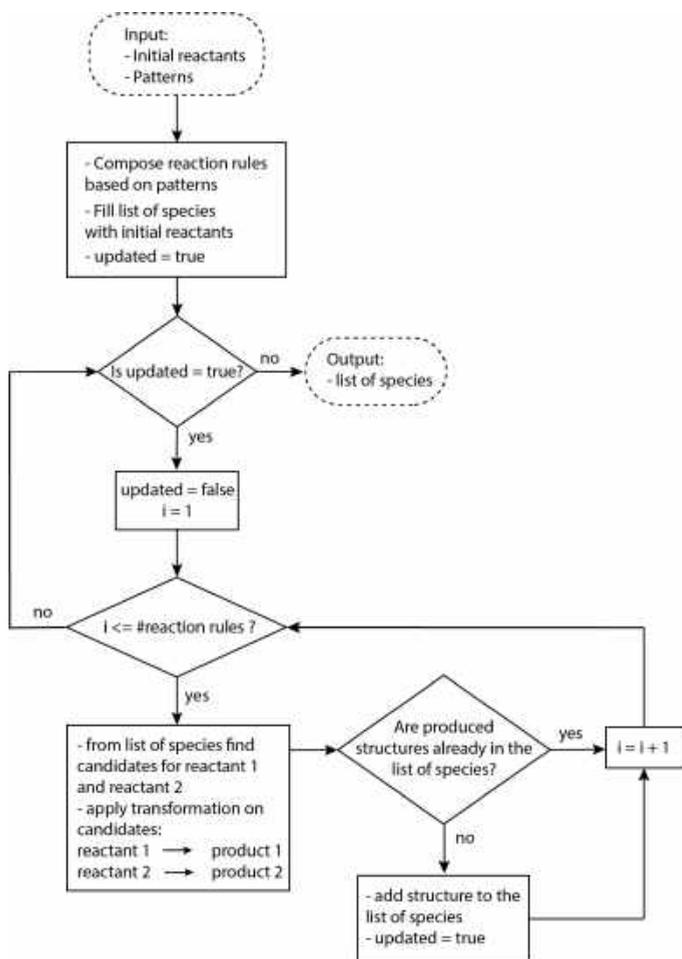
2

Figure 2.5: Main steps of the reaction generating mechanism.

After the transformation is applied to the candidate species, a product species is generated, which might be already featured in the list of species. In that case the algorithm proceeds with the next reaction, but if the molecular structure is new, then it is added to the list. To check whether a molecule is already in the list, the Ullmann procedure is applied again. In this case, the isomorphism problem for a whole molecule graph rather than for a subgraph has to be solved: the whole molecular structure has to be compared to other molecules and an exact match for all its atoms has to be found. This implies a modification of the matching criteria. The degree condition is strict: nodes are matched only if their degrees are the same. Furthermore, the crosslink condition is bijective: nodes are matched only if both carry a crosslink of the same type.

One iteration of the procedure is completed when all reaction rules are visited. The next iteration is performed if at least one new species is created during the previous iteration. Newly created species may still undergo some reactions from the list of reaction rules. The procedure stops if the reaction rules fail to extend the list of species. As a result, the procedure generates all intermediate and final molecular structures which are produced in a system due to specified reaction rules. The whole algorithm is illustrated in the flowchart of Figure 2.5.

2.3.2. Reaction network

Every time a reaction rule "fires", identifiers of reactant and product species together with the type of reaction are stored. Identifiers of molecular species correspond to their ordinal numbering in the list of species the construction of which was described in subsection 2.3.1.

Having information about the detailed reaction mechanism, one can reconstruct the reaction network. The reaction network is a systematic representation of the reaction mechanism. Let's assume, that the output of the reaction generating mechanism produced n different species, which took part in m reactions. The reaction network is a directed labeled bipartite graph $G = (l, V, E)$ of size $n + m$ with two types of nodes $V = \{V_S, V_R\}$: species and reactions. Nodes V_S have a weight corresponding to the initial concentration c_0 of a species in the system. The weight of nodes v_R denotes the rate coefficient k of the reaction.

Concerning directionality, each node of the reaction network has incoming, outgoing, or both types of edges. Let us assume, two species v_{S1} and v_{S2} take part in a second order reaction v_R and produce species v_{S3} and v_{S4}. This reaction is illustrated in Figure 2.6 (a). Both reactant species v_{S1} and v_{S2} have outgoing edges which point to the reaction node v_R. The reaction node has two outgoing edges pointing toward each of

the product species, v_{S3} and v_{S4}. Furthermore, reactant and product species can take part in other reactions and have more outgoing and incoming edges.

Species nodes are unique: each molecular structure is represented by a node which appears only once in the whole reaction network. Reaction nodes are labeled according to the reaction rule that created the reaction. Thus, the reaction nodes can be partitioned into families of reactions.

Initial species nodes do not have incoming edges. Species nodes having both incoming and outgoing edges take part in intermediate reactions. Species which have only incoming edges correspond to the final products of the reaction mechanism. As for the reaction nodes, their incoming and outgoing degree is determined by the order of a reaction and number of product species they produced.

2.3.3. Kinetic model from reaction network

An important goal of kinetic polymer modeling is to compute the concentration of all species over time given the set of reactions and kinetic coefficients. This is usually performed manually by setting up ordinary differential equations, describing the mass balances of all species, from a given set of reaction equations. However, the availability of the reaction network as described in the previous paragraph enables an automated procedure to construct the ODE's, since the network carries all information sufficient for kinetic model construction. The mass balance equations can be constructed for each species by exploring its first and second order neighborhood in the reaction network. All of these equations contain contribution and subtraction terms. A contribution term originates from a reaction where the species acts as a product. A subtraction term denotes the species as a reactant. As an example Figure 2.6 (b) shows fragments of the inimer/isobutylene copolymerization reaction network, which illustrates how the contribution and the subtraction terms are constructed from the environment of species node v_S.

In the following section we discuss how the adjacency matrix A and weights of the nodes of the reaction network are transformed into the kinetic model of the system. Consider a reaction network with reactions of type $A \to B_1 + B_2 + \dots$ and $A + B \to B_1 + B_2 + \dots$ The main assumption here is that the order of a reaction is equal to the number of reactants: consequently, only first and second order reactions are considered.

- **First order reactions**

If the reaction node i represents a first order reaction then the in-degree of this node is

one, and the out-degree is non-zero. That is

$$\deg_{\text{in}}(i) := \sum_{j=1}^{n} A_{j,i} = 1, \deg_{\text{out}}(i) := \sum_{j=1}^{n} A_{i,j} > 0.$$

For a species node i, the differential equation describing the evolution of species concentration $c_i(t)$ is written in terms of the adjacency matrix A and coefficient vector k,

$$c_i'(t) = \sum_{j=1}^{n} A_{j,i} k_j \sum_{l=1}^{n} A_{l,j} c_l(t) - c_i(t) \sum_{j=1}^{n} A_{i,j} k_j \tag{2.1}$$

$$= \sum_{j=1}^{n} A_{j,i} k_j \sum_{l=1}^{n} A_{l,j} c_l(t) - \sum_{j=1}^{n} A_{i,j} k_j \sum_{l=1}^{n} A_{l,j} c_l(t)$$

$$= \sum_{j=1}^{n} k_j (A_{j,i} - A_{i,j}) \sum_{l=1}^{n} A_{l,j} c_l(t).$$

When rewritten in terms of a matrix multiplication, it becomes clear that these equations constitute an autonomous linear system of ODEs,

$$\frac{\mathrm{d}}{\mathrm{d}t} c(t) = \left(A^T \text{diag}\{k\} A^T - \text{diag}\{Ak\} \right) c(t). \tag{2.2}$$

Let, $c(0) = c_0$, then the formal solution to Eq. (2.2) is expressed via the matrix exponentiation,

$$c(t) = e^{\left(A^T \text{diag}\{k\} A^T - \text{diag}\{Ak\} \right) t} c_0.$$

- **Second order reactions**

Equation (2.1) can be extended to cover reactions of arbitrary order,

$$c_i'(t) = \sum_{j=1}^{n} A_{j,i} k_j \prod_{l=1}^{n} c_l^{A_{l,j}} - \sum_{j=1}^{n} A_{i,j} k_j \prod_{l=1}^{n} c_l^{A_{l,j}} = \tag{2.3}$$

$$\sum_{j=1}^{n} k_j (A_{j,i} - A_{i,j}) \prod_{l=1}^{n} c_l^{A_{l,j}}.$$

For a fixed reaction node j, consider a matrix with elements $S_{i,l} = \frac{1}{2} A_{i,j} A_{l,j}$ if $i \neq l$ and $S_{i,l} = 0$, if $i = l$. If i, l are the reactant species for a second order reaction j then $S_{i,l} = \frac{1}{2}$. Furthermore, if node j represents a second order reaction, then the in-degree for this node is two, $\deg_{\text{in}}(j) = 2$, and the product appearing in Eq. (2.3) contains only two terms, $\prod_{l=1}^{n} c_l^{A_{l,j}} = c^T S c = \frac{1}{2} c^T A_{\cdot,j} A_{\cdot,j}^T c$, where $A_{\cdot,j}$ is the jth column of matrix
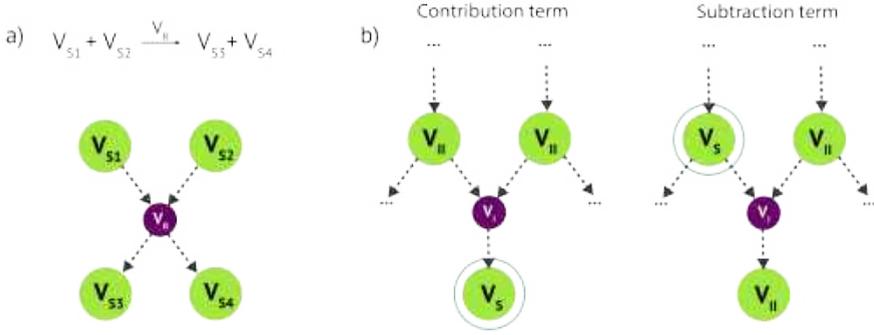
Figure 2.6: (a) A fragment of the reaction network, which illustrates the general case of a second order reaction, where $V_{Si}, i = 1 \ldots 4$ nodes are species which are taking part in the reaction and V_R represents the type of the reaction rule. (b) Nodes influencing the concentration of species V_S. On the left hand side, the two different second order neighbors V_{II} of V_S augment its concentration due to the reaction V_I, which is represented by the first order neighbor from the incoming edge. On the right hand side, the concentration of V_S is decreasing by undergoing the reaction as denoted by V_I.

A. Thus, Eq. (2.3) simplifies to

$$c'_i(t) = c^T(t) M_i\, c(t), \tag{2.4}$$

where $M_i = \sum_{j=1}^{n} \frac{1}{2} k_j A_{\cdot,j} A_{\cdot,j}^T (A_{j,i} - A_{i,j})$.

- **Mix of first and second order reactions**

Let us split the vector of reaction rates $k = k^{(1)} + k^{(2)}$, so that $k^{(1)}$ is non-zero for first order reactions and $k^{(2)}$ for second order ones. A combination of Eqs. (2.2) and (2.4), together with the corresponding rate vectors $k^{(1)}, k^{(2)}$ yields the differential equation for a mixed reaction network,

$$c'_i(t) = c(t)^T M_i\, c(t) + B_{i,\cdot} c(t), \tag{2.5}$$

where

$$M_i = \sum_{j=1}^{n} \frac{1}{2} k_j^{(2)} A_{\cdot,j} A_{\cdot,j}^T (A_{j,i} - A_{i,j})$$

and

$$B = A^T \text{diag}\{k^{(1)}\} A^T - \text{diag}\{A k^{(1)}\}.$$

Note, $A_{\cdot,j} A_{\cdot,j}^T$ is an outer product (column times row).

Ultimately, solving Eqs. (2.2), (2.4) and (2.5) provides information about the distribution of the concentrations of all species present in the resulting reaction network.
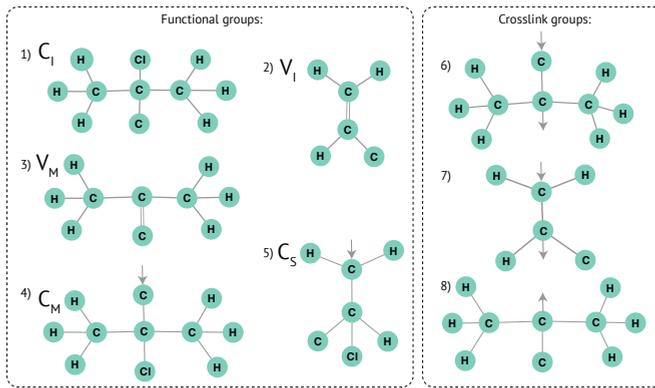
Figure 2.7: Complete set of patterns to describe the reaction rules of IB/IM copolymerization (Table 2.1). The five reactive functional groups are each represented by a pattern: C_I, C_M, C_S, V_I, V_M. Another 3 patterns are required to represent different crosslinked species.

Table 2.2: Reaction steps for copolymerization of IM with IB in terms of patterns

#	Reaction	Rate coefficient ($Lmol^{-1}s^{-1}$)
1	pattern 1 + pattern 2 → pattern 8 + pattern 5	3.32×10^{-2}
2	pattern 1 + pattern 3 → pattern 8 + pattern 4	4.46×10^{-4}
3	pattern 4 + pattern 2 → pattern 6 + pattern 5	5.19×10^{-1}
4	pattern 4 + pattern 3 → pattern 6 + pattern 4	2.27×10^{0}
5	pattern 5 + pattern 2 → pattern 7 + pattern 5	6.45×10^{-3}
6	pattern 5 + pattern 3 → pattern 7 + pattern 4	4.11×10^{-5}

Moreover, a mix of first and second order reactions is the most common case for polymer reaction engineering problems.

2.4. From reaction network to macromolecular properties

In this section an application of the reaction generating mechanism is illustrated on the copolymerization of inimer (IM) with isobutylene (IB) in a batch reactor. In Table 2.1 reaction steps between IM and IB are shown together with the rate constants found in the literature (Zhao et al. (2016), Zhao et al. (2013)). The objective of this section is to demonstrate how we arrive at macromolecular properties for a polymerization using reaction generating mechanism described in the sections above.
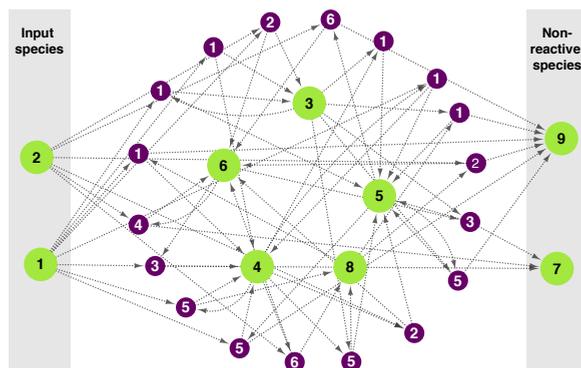
Figure 2.8: The reaction network after applying the reaction generating mechanism for copolymerization of IM and IB. Nine green nodes correspond to species and 18 purple nodes, with numbering corresponding to the list of reactions in Table 2.1, denote reactions. Species node 9 contains the IM product patterns and node 7 contains the IB product patterns shown at the right hand side in Figure 2.7.

Firstly, IM and IB are converted to molecular graphs and the reaction rules are constructed based on the reaction mechanism described in Table 2.1. A reaction system can only be fully described, if a complete set of reaction rules is available, hence we need a complete set of patterns as well. Such a complete set is obtained by inspection of all the structural formula's concerned with the initial and final species of the reactions involved and all possible intermediates. For the IM/IB copolymerization system a complete set of eight patterns is obtained, as shown in Figure 2.7, five representing the reactive groups C_I, C_M, C_S, V_I, V_M, and three more final product patterns all having crosslinks. The reaction steps from Table 2.1 are represented as the reaction rules on patterns in Table 2.2.

Only directional crosslinks are present in the system, as all of the crosslinking reactions happen due to the chlorine atom reacting with the vinyl group. IB may have at most one incoming and one outgoing crosslink at the same time. The incoming crosslink is formed due to the chlorine attacking the vinyl group of IB. After this reaction, the chlorine is transferred to IB and propagates the polymer chain, thus IB gets the outgoing crosslink. The IM monomer has a vinyl group on one side and the chlorine on the other side of it. The chlorine thus attacks any vinyl group in the system and forms an outgoing crosslink. Meanwhile, the vinyl group on the other side of the molecule reacts with the chlorine and forms an incoming crosslink. Further, the chlorine is transferred to the former vinyl side of IB and undergoes another crosslinking reaction. In this configuration, IM may have at most one incoming and two outgoing crosslinks.
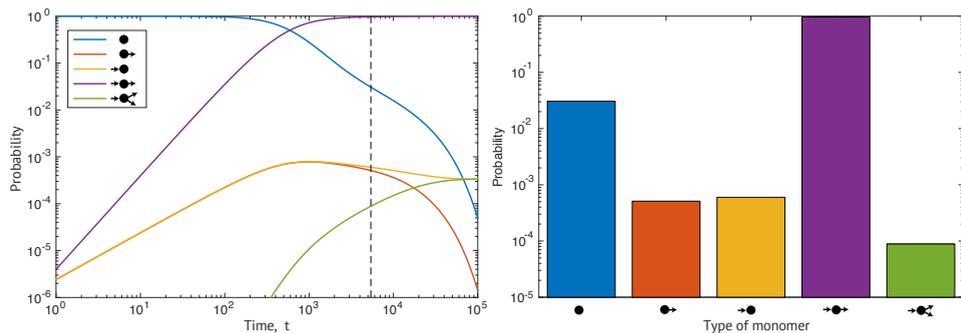
Figure 2.9: Probability profile (normalized concentration) over time of monomers having different configuration of crosslinks in a batch reactor, starting with initial IB and IM concentrations $c_{IB} = 1.74$ $mol \cdot L^{-1}$, $c_{IM} = 0.000454$ $mol \cdot L^{-1}$ [33]. The vertical line at $t = 5400$ s marks the concentration distribution or degree distribution that is used to compute the number distribution of monomer units

Having prepared all the initial species, patterns and reaction rules the algorithm is applied to the example of IB/IM copolymerization. As a result, seven more molecular structures are obtained. We thus generated the complete reaction network as shown in Figure 2.8. The system with eight patterns turns out to yield a total of eighteen reactions between nine species.

Furthermore, the mass balance equations to the IB/IM copolymerization were generated from the reaction network using equations described in Section 3.2. The ODEs are solved for initial IB and IM concentrations $c_{IB} = 1.74$ $mol \cdot L^{-1}$, $c_{IM} = 0.000454$ $mol \cdot L^{-1}$ [33]. This gives the time profiles of all the crosslinked species until $t = 10^5$ s as shown in Figure 2.9. This in fact represents the time profile of the *degree distribution* forming the starting point for the derivation of macromolecular properties using random graph theory as will be discussed in the next paragraph. All calculations were performed using single core processor. The reaction network construction was completed in 26 seconds and the ODEs were solved in less than 2 seconds.

The degree distribution is obtained by a normalisation of instantaneous concentration of cross-linked species:

$$u(k,l) = \frac{c_{k,l}(t)}{\sum\limits_{k,l=0}^{\infty} c_{k,l}(t)} \tag{2.6}$$

where $c_{k,l}(t)$ is the molar concentration of a crosslinked species with k incoming crosslinks and l outgoing at the time t. Thus, $u(k,l)$ is a probability distribution that describes the local connectivity of monomer units at a given time .

Random graph theory is then used to obtain average properties for the set of polymer
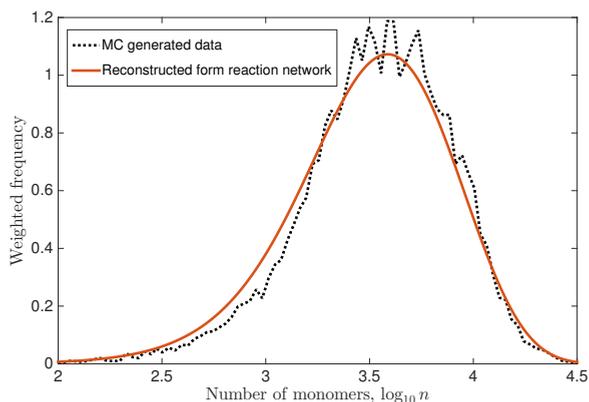
2

Figure 2.10: Double weighted molecular weight distribution for IB/IM copolymerization in a batch reactor with initial concentrations $c_{IB} = 1.74 \ mol \cdot L^{-1}$, $c_{IM} = 0.000454 \ mol \cdot L^{-1}$ at $t = 5400$ s compared with Monte Carlo generated data from[33]

networks all of which comply with the same degree distribution[38]. The random graph theory has been successfully applied to a number of branched polymer systems, either or not forming gels, enabling to compute macromolecular properties like size distributions. The details of the mathematical procedures to obtain the number distribution of monomers of the IB/IM copolymerization from the degree distribution (Figure 2.9) are given in the Appendix A1. The distribution thus obtained, for a batch time of $t = 5400$ s (vertical line in Figure 2.9), is shown in Figure 2.10, where it is compared to the Monte Carlo simulation results by[33]. One observes a very good agreement. It should be realized that this is the first deterministic solution of a full probability distribution of numbers of monomers for IB/IM copolymerization. Moreover, in terms of classical polymerization modeling this is also a very challenging mathematical problem as it requires simultaneous solving of number distributions for several reactive groups C_M, C_S, V_M, V_I.

2.5. Conclusions

We have designed a new methodology for the prediction of macromolecular properties of polymerization systems possessing a kinetic complexity that is comparable to gas phase pyrolysis of hydrocarbons and biochemical reactions. Such complex systems have justified the development of automated reaction generation tools to cover thousands of species and reactions, far more than polymer systems that usually are described with simpler or simplified kinetic schemes. The direct application of available automated re-

action tools is prohibited by the sheer size of the polymer molecules, eventually growing to infinity. Representing polymers of a reasonable size as molecular graphs is virtually impossible. Therefore, we decided to design a novel automated polymer reaction methodology based on a *monomer approach*.

Existing tools employ molecular graphs. We utilize molecular graphs as well to describe monomer units, but as different from the existing tools, we add features to these graphs that denote directional crosslinks. Thus, our methodology proposes a compromise: employing molecular graphs of acceptable size we still can estimate some average properties of the polymer network.

We have proven the success of our methodology using the example of copolymerization of inimer and isobutylene. Starting with the molecular graphs of IM and IB the methodology proceeds through the formulation of reaction rules and patterns describing the copolymerization, the automatic generation of a reaction network and construction of mass balance equations. Solution of the latter provides the probability distribution of local monomer connectivity - the degree distribution. By using random graph theory this leads to a molecular weight distribution that coincides with results from Monte Carlo simulations.

Essentially, in this manner we demonstrated how to proceed from a kinetic scheme to macromolecular properties in an automated manner. It should be realized that the example taken still has relatively simple kinetics, but our methodology allows extension to really complex kinetics in a straightforward manner, as we will show in future studies.

References

[1] M. Nijemeisland, L. K. E. A. Abdelmohsen, W. T. S. Huck, D. A. Wilson, and J. C. M. van Hest, *A compartmentalized out–of–equilibrium enzymatic reaction network for sustained autonomous movement,* ACS Cent. Sci. **2**, 843 (2016).

[2] D. Y. Zubarev, D. Rappoport, and A. Aspuru-Guzik, *Uncertainty of prebiotic scenarios: The case of the non–enzymatic reverse tricarboxylic acid cycle,* Scientific Reports **5** (2015), 10.1038/srep08009.

[3] H. Jeong, B. Tombor, R. A. Z. N. Oltvai, and A.-L. Barabási, *The large-scale organization of metabolic networks,* Nature **407**, 651 (2000).

[4] J. Stelling, S. Klamt1, K. Bettenbrock, S. Schuster, and E. D. Gilles, *Metabolic network structure determines key aspects of functionality and regulation,* Nature **420**, 190 (2002).

[5] I. Otero-Muras, P. Yordanov, and J. Stelling, *Chemical reaction network theory elucidates sources of multistability in interferon signaling,* PLOS Com. Boil. **13** (2017).

[6] I. Kryven, S. Röblitz, and C. Schütte, *Solution of the chemical master equation by radial basis functions approximation with interface tracking,* BMC Systems Biology **9** (2015), 10.1186/s12918-015-0210-y.

[7] E. Lounkine, M. J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J. L. Jenkins, P. Lavan, E. Weber, A. K. Doak, S. Côté, B. K. Shoichet, and L. Urban, *Large-scale prediction and testing of drug activity on side-effect targets,* Nature **486**, 361 (2012).

[8] C. Palaniappan, R. Srinivasan, and R. Tan, *Expert system for the design of inherently safer processes. 1. route selection stage,* Ind. Eng. Chem. Res. **41**, 6698 (2002).

[9] G. Shinar and M. Feinberg, *Structural sources of robustness in biochemical reaction networks,* Science **5971**, 1389 (2010).

[10] E. Ranzi, T. Faravelli, P. Gaffuri, and A. Sogaro, *Low-temperature combustion: automatic generation of primary oxidation reactions and lumping procedures,* Combustion and Flame **102**, 179 (1995).

[11] F. Battin-Leclerc, *Development of kinetic models for the formation and degradation of unsaturated hydrocarbons at high temperature,* Physical Chemistry Chemical Physics **4**, 2072 (2002).

[12] L. J. Broadbelt and J. Pfaendtner, *Lexicography of kinetic modeling of complex reaction networks,* AIChE J **51**, 2112 (2005).

[13] J. Song, *Building robust chemical reaction mechanisms : next generation of automatic model construction software*, Ph.D. thesis, Massachusetts Institute of Technology (2004).

[14] N. M. Vandewiele, K. M. V. Geem, M.-F. Reyniers, and G. B. Marin, *Genesys: Kinetic model construction using chemo-informatics,* Chemical Engineering Journal **207–208**, 526 (2012).

[15] S. Rangarajan, A. Bhan, and P. Daoutidis, *Language-oriented rule-based reaction network generation and analysis: Description of RING,* Computers and Chemical Engineering **45**, 114 (2012).

[16] P. H. M. V. Steenberge, D. R. D'hooge, Y. Wang, M. Zhong, M.-F. Reyniers, D. Konkolewicz, K. Matyjaszewski, and G. B. Marin, *Linear gradient quality of atrp copolymers,* Macromolecules **45**, 8519–8531 (2012).

[17] D. R. D'hooge, P. H. M. V. Steenberge, P. Derboven, M.-F. Reyniers, and G. B. Marin, *Model-based design of the polymer microstructure: bridging the gap between polymer chemistry and engineering,* Polymer Chemistry **6**, 7081 (2015).

[18] P. H. M. V. Steenberge, B. Verbraeken, M.-F. Reyniers, R. Hoogenboom, and D. R. D'hooge, *Model-based visualization and understanding of monomer sequence formation in gradient copoly(2-oxazoline)s on the basis of 2-methyl-2-oxazoline and 2-phenyl-2-oxazoline,* Macromolecules **48**, 7765– (2015).

[19] T. Lemos, P. A. Melo, and J. C. Pinto, *Stochastic modeling of polymer microstructure from residence time distribution,* Macromolecular Reaction Engineering **9**, 259 (2015).

[20] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness* (Freeman, New-York, 1979).

[21] H. Tobita and A. E. Hamielec, *Polymerization processes, 2. modeling of processes and reactors,* in *Ullmann's Encyclopedia of Industrial Chemistry* (Wiley-VCH Verlag GmbH & Co. KGaA, 2000).

[22] P. Iedema and N. Kolhapure, *Mathematical methods,* in *Handbook of Polymer Reaction Engineering* (Wiley-VCH Verlag GmbH, 2008) pp. 431–532.

[23] P. A. Mueller, J. R. Richards, and J. P. Congalidis, *Polymerization reactor modeling in industry,* Macromolecular Reaction Engineering **5**, 261 (2011).

[24] D. R. D'hooge, P. H. Steenberge, M.-F. Reyniers, and G. B.Marin, *The strength of multi-scale modeling to unveil the complexity of radical polymerization,* Progress in Polymer Science **58**, 59 (2016).

[25] E. Mastan, X. Li, and S. Zhu, *Modeling and theoretical development in controlled radical polymerization,* Progress in Polymer Science **45**, 71 (2015).

[26] L. H. Oakley, F. Casadio, K. R. Shull, and L. J. Broadbelt, *Microkinetic modeling of the autoxidative curing of an alkyd and oil-based paint model system,* Applied Physics A: Materials Science and Processing **121**, 869 (2015).

2

[27] P. Iedema, J. Hermans, K. Keune, A. van Loon, and M. Stols-Witlox, *Mathematical modeling of mature oil paint networks*. in *ICOM-CC 17th Triennial Conference Preprints*, edited by J. Bridgland (Paris: International Council of Museums, Melbourne, 2014.) p. 1604, iSBN 978-92-9012-410-8.

[28] I. Kryven, *Analytic results on the polymerisation random graph model,* Journal of Mathematical Chemistry (2017), 10.1007/s10910-017-0785-1.

[29] I. Kryven, *General expression for the component size distribution in infinite configuration networks,* Phys. Rev. E **95**, 052303 (2017).

[30] V. Schamboeck, I. Kryven, and P. D. Iedema, *Acrylate network formation by free-radical polymerization modeled using random graphs,* Macromolecular Theory and Simulations , 1700047 (2017).

[31] I. Kryven, J. Duivenvoorden, J. Hermans, and P. D. Iedema, *Random graph approach to multifunctional molecular networks,* Macromolecular Theory and Simulations **25**, 449– (2016).

[32] J. E. Puskas and M. Grasmüller, *Star-branched and hyperbranched polyisobutylenes,* Macromolecular Symposia **132**, 117 (1998).

[33] Y. R. Zhao, K. B. McAuley, P. D. Iedema, and J. E. Puskas, *Advanced Monte Carlo modeling using weight-based selection of arborescent polyisobutylene molecules in a batch reactor,* Macromolecular Theory and Simulations **25**, 134 (2016).

[34] A. Bondy and M. R. Murty, *Graph Theory* (Springer-Verlag London, 2008).

[35] J. Lee, W.-S. Han, R. Kasperovics, and J.-H. Lee, *An in-depth comparison of subgraph isomorphism algorithms in graph databases,* Proceedings of the VLDB Endowment **6**, 133 (2012).

[36] J. Ullmann, *An algorithm for subgraph isomorphism,* Journal of the Assiciation for Computing Machinery **23**, 31 (1976).

[37] K. Gouda and M. Hassaan, *A fast algorithm for subgraph search problem,* in *Proc. of the 8th International Conference on Informatics and Systems* (2012) p. DE–53 – DE–59.

[38] I. Kryven, *Emergence of the giant weak component in directed random graphs with arbitrary degree distributions,* Physical Review E **94**, 012315 (2016).