Karlsruhe Institute of Technology

Classification of Gas Networks: A Graph-Theoretic Approach

Master Thesis of

Marc Jenne

At the Department of Informatics
Institute of Theoretical Informatics

Reviewers:	PD Dr. Torsten Ueckerdt
	Jun.-Prof. Dr. Thomas Bläsius
Advisors:	PD Dr. Torsten Ueckerdt
	Sascha Gritzbach
	Matthias Wolf

Time Period: 1st June 2022 – 1st December 2022

Acknowledgements

I would like to thank Louis Wayas of the DVGW EBI for providing valuable knowledge about gas networks and for providing a data set of gas network instances.

Statement of Authorship

Ich versichere wahrheitsgemäß, die Arbeit selbstständig verfasst, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde sowie die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, December 1, 2022

Abstract

Gas networks account for a large share of energy supply both in households and industry. They are present in all region types, such as large cities, rural areas and industrial areas. It is of interest to see whether gas networks can be classified based on the region type which they originate from. This is possible if gas networks from similar regions share important common characteristics and if gas networks from different regions differ significantly from each other.

We present a graph-theoretic approach to characterise and classify gas networks originating from different region types. For this purpose, we model gas networks as graphs and examine several parameters that describe graphs from various perspectives. Using a data set of gas network instances with known origins, we analyse the common features and differences both between networks from the same region type and between networks from different types. Based on these analyses we find distinguishable gas network classes as well as meaningful parameters that characterise these classes.
We then construct gas network classifiers based on three different classification approaches. Each of these classifiers determines which region type a gas network instance with unknown origin most likely originates from. We evaluate the different classification approaches, compare them and discuss their strengths and weaknesses.

Deutsche Zusammenfassung

Gasnetze haben sowohl in Haushalten als auch in der Industrie einen großen Anteil an der Energieversorgung. Sie sind in allen Arten von Regionen vorhanden, wie zum Beispiel in Großstädten, ländlichen Gebieten und Industriegebieten. Eine interessante Fragestellung ist, ob sich Gasnetze anhand der Art der Region, aus der sie stammen, klassifizieren lassen. Das ist genau dann möglich, wenn Gasnetze aus ähnlichen Regionen wichtige gemeinsame Merkmale aufweisen und wenn sich Gasnetze aus verschiedenen Regionen deutlich voneinander unterscheiden.

Wir stellen einen graphentheoretischen Ansatz zur Charakterisierung und Klassifizierung von Gasnetzen vor, die aus verschiedenen Arten von Regionen stammen. Dafür modellieren wir Gasnetze als Graphen und untersuchen verschiedene Parameter, die Graphen aus unterschiedlichen Perspektiven beschreiben. Anhand eines Datensatzes von Gasnetzinstanzen mit bekannter Herkunft analysieren wir die Gemeinsamkeiten und Unterschiede sowohl zwischen Netzen aus derselben Region als auch zwischen Netzen aus verschiedenen Regionen. Mit diesen Analysen finden wir unterscheidbare Gasnetzklassen sowie aussagekräftige Parameter, die diese Klassen charakterisieren.
Anschließend konstruieren wir Gasnetzklassifikatoren, die auf drei verschiedenen Klassifizierungsansätzen basieren. Jeder dieser Klassifikatoren bestimmt, aus welcher Art von Region eine Gasnetzinstanz mit unbekannter Herkunft am wahrscheinlichsten stammt. Wir bewerten die verschiedenen Klassifizierungsansätze, vergleichen sie und diskutieren ihre Stärken und Schwächen.

Contents

1. Introduction

Natural gas networks are one of the most important infrastructures in Europe. In 2021, around 40 % of households are connected to the European gas network that consists of over 2,000,000 km of pipelines. Natural gas represented 21.5 % of the primary energy consumption in the European Union and held the largest share of energy supply in households [GEU].

The importance of gas networks makes them an interesting research topic. It stands to reason that the characteristics of gas networks vary between different region types (hereafter often simply referred to as regions), such as inner cities on the one hand and industrial areas on the other hand. This assumption raises two important questions. First, it is of interest whether gas networks in similar regions share the same characteristics, by which they can therefore be identified. The second arising question is whether gas networks in different regions differ significantly from each other. If both of these questions can be answered in the affirmative, it is possible to classify gas networks into different classes based on the region which they originate from. Furthermore, each class can be described by the characteristics that all networks of that class have in common and that separate them from the other classes.

With the knowledge of the existing gas network classes and their features, a gas network instance with unknown origin can be assigned a region label, i.e., one can compute the region type the instance originates from.

Having detailed insights into the characteristics of gas networks in a specific region also opens further applications, namely the generation of generic gas network instances. Such an instance captures the most important characteristics of the region class it belongs to. The need for generic networks arises both from the objective of examining the current state of gas networks and from carrying out simulations of the future development of gas networks. When an actual gas network of a specific region is to be examined, modelling that network may be very time-consuming or not possible at all, for example because important information is not available. In these cases, one can generate a generic network of similar type and size and examine this network instead of the actual one.
Another possible field of application is the evolution of regions. When the characteristics of a region are expected to change, the impact of these local gas network changes on the larger overall gas network it is part of can be simulated. Two possible scenarios are, for example, a village that is expected to grow into a city within the next few years, or a new industrial area that is planned to be constructed in place of a given residential area. With

generic gas networks of a city or an industrial area these changes can be simulated. In this way, for instance, bottlenecks in the overall network may be detected and avoided. Moreover, the generic networks may help to estimate resource demand, construction cost and similar key metrics for the changes in the regional networks.
Another possible scenario in the evolution of regions is a decline in gas demand. Take, for example, a city that is expected to neither grow nor shrink, but a significant portion of the consumers are expected to switch to renewable energies. In this scenario, the existing gas network has to be replaced by a network with similar characteristics, but of smaller size. Again, with generic networks various simulations can be performed providing helpful information for planning.

1.1. Related Work

The field of analysing and comparing the characteristics of gas networks has not been intensively investigated. Especially focussing on the examination and classification of regional gas networks, no such studies have been done yet. In 2022, Ye et al. [YLLL22] analysed the topology of a Chinese and a European network based on complex network theory. Among other results, they found higher redundancy and greater robustness in the European network compared to the Chinese one. However, while they examined partly similar characteristics as we aim for, their networks and results are not comparable to the regional networks we are interested in, as the former cover a much larger area and are modelled on a different level of detail.
In 2020, Then et al. [TSB+20] addressed the problem of modelling a decline in gas consumption in urban gas networks. For this purpose, they examined the relationship between network length and customer amount. They found a relationship described by a power law with an exponent that is correlated with several topological parameters of the gas network. With these findings they showed that the decline in gas network costs is much slower than the corresponding decline in demand. They also stated that more detailed insights into the topological parameters may be helpful for more accurate cost estimates.

A field closely related to gas networks are street networks, as gas pipelines are often laid underneath streets. Therefore, both networks often share their topological features. In comparison to gas networks, street networks have been investigated more. Still, actual comparisons regarding the characteristics of street networks in different region types have not been investigated yet. Instead, most works focus solely on one specific region. In 2019, Zhu et al. [ZSG19] examined rural traffic networks with regard to their topological characteristics and their vulnerabilities. In 2017, Lin and Ban [LB17] performed a comparative analysis on topological structures of three urban street networks, namely Stockholm, Toronto and Nanjing. In 2019, Sharifi [Sha19] studied urban street networks with a focus on the influence of street networks on the urban resilience.

Other networks related to gas networks are water networks and electrical grids, as the way they are laid is quite similar. However, no actual comparisons between network characteristics across different regions exist in these fields either. For urban water networks, Krueger et al. [KKU+17] performed an analysis of generic patterns in an Asian city in 2017. Studies related to the topology of electrical distribution grids in rural areas and urban areas are reported in [RFRW22] and [LWR16], respectively.

1.2. Contribution

We present a graph-theoretic approach to characterise and classify gas networks. Our focus is on regional gas networks, i.e., networks that cover areas like a city or a rural area. The networks we work with come from a data set provided by the DVGW Research

Center at the Engler-Bunte-Institute of Karlsruhe Institute of Technology [DRC]. This data set consists of gas networks from Germany, each of them labelled with the region type it originates from. These types are inner cities, old towns, residential areas, rural areas and industrial areas. Hence, in this work we deal with these five regions. However, our methodology can be applied to any data set and set of regions.
The problem we want to solve is whether and how gas networks originating from different region types can be clearly distinguished from each other. For this purpose, we model gas networks as graphs. We then identify several graph parameters that describe these graphs in different ways. Using these parameters, we analyse the gas network instances of our data set. In that analysis, for each parameter we compare the gas networks among each other to examine differences and common features both between the instances of the same region and between the different regions. We identify parameters that are meaningful to describe the characteristics of the networks of certain regions and to distinguish between networks of different regions. With the found results we evaluate whether a classification into distinct classes based on the regions is possible, or whether the networks across all regions are too similar to each other to distinguish between them.
As mentioned previously, in this work we focus on regional gas networks. In addition, we also provide a brief comparison between regional and supraregional networks. The latter are gas networks that cover larger areas and operate on a much higher pressure.

With the set of meaningful parameters and classes found, we then present three classifier approaches for the gas network classification problem. These classifiers assign a class label to a gas network instance, i.e., compute which region the instance most probably originates from. We evaluate the different approaches and discuss their strengths and weaknesses. Furthermore, we provide an outlook on how these classifiers can be improved in the future.

1.3. Outline

In Chapter 2, we introduce the basics of graph theory as well as the field of classification problems. Furthermore, we present parameters from descriptive statistics we use later.
Next, in Chapter 3, we explain how we model gas networks as graphs and formally state the problem we solve in this work. We also describe the different gas network regions we examine and provide some expectations we have regarding the characteristics of these regions.
In Chapter 4, we present several graph parameters that describe a graph from various perspectives. These parameters are then analysed in Chapter 5. At the end of this chapter, we state our overall results regarding the questions whether different gas network classes exist and which parameters we consider meaningful to characterise and identify these classes. In addition, we also compare the instances analysed so far with a supraregional network, though the latter is not considered an additional class.
With these findings, in Chapter 6 and Chapter 7 we first construct and then evaluate and discuss three gas network classification approaches. We also provide an outlook on how these classifiers can be adjusted and refined with more data available.
Lastly, in Chapter 8, we summarise our results and provide an outlook on further research topics.

2. Preliminaries

In this chapter, we will introduce the basic aspects and notations of graph theory. Furthermore, we provide an introduction into the field of classification und clustering problems and present one specific classifier. Lastly, we explain some frequently used parameters from descriptive statistics.

2.1. Graphs

A graph G is an abstract representation of objects and relations between those objects. Formally, a graph $G = (V, E)$ is a tuple consisting of a set of nodes V, representing the objects, and a set of edges E. Each edge $e \in E$ represents the relation between two nodes u and v. It can be either *directed* or *undirected*; in the first case, $e = (u, v)$ is an ordered pair and we call u the start node and v the end node of e. In the latter case, $e = \{u, v\}$ is an unordered pair and we call both u and v the end nodes of e. In both cases, we call u a neighbour of v and vice versa. An edge is called *loop* if both of its end nodes are the same node.
Depending on the application, one or multiple attributes can be associated with the nodes and edges. Probably the most common example for such an attribute is the *weight* or *length* of an edge.
A path $p = (u_1, u_2, ..., u_n)$ in a graph is an ordered list of nodes where each edge $\{u_i, u_{i+1}\} \in E$. The length of a path is defined as the sum of all edge lengths on the path. We call an undirected graph G *connected* if for each unordered pair of nodes $\{u, v\}$ a path leading from u to v exists.
Another important attribute of nodes is the *node degree*. For an undirected graph and a node $u \in V$, its node degree $d(u)$ is the number of edges incident to u:

$$d(u) = |\{e = \{x, y\} \in E \mid x = u \text{ or } y = u\}|. \tag{2.1}$$

The node degree for a directed graph is defined analogously, but we can furthermore distinguish between the *in-degree* and the *out-degree* of nodes.

Graphs in this work are assumed to be *simple* graphs unless stated otherwise. In such graphs at most one edge exists between any pair of nodes. In contrast, there are also *multigraphs* where multiple edges between two nodes can exist. Each of these edges can have different attributes or attribute values.

2.2. Classification

Our overall goal in this work is to assign category labels to gas networks, based on the characteristics of the network. This class of problems is generally known as Classification Problems. In this section, we formally define this problem and introduce an example of a classifier as well as known problems we may encounter. Moreover, we provide a short explanation of the related Clustering Problem and how it differs from classification. More detailed introductions into classification and clustering are provided in the studies of [Dou12] and [Agg18]. These works also comprise examples of specific classification and clustering approaches.

2.2.1. The Classification Problem

Classification is the problem of assigning a category, called class, to each instance of a data set D. It is a *supervised* learning task, which means that the set of classes Y is known upfront. Additionally, a training set $T \subset D$ of *labelled* instances is available as a set of instances for which the class label is known beforehand.
The assignment of class labels to instances is done by a *classifier*. Most classifiers are built up in a training phase. In that phase, the classifier uses the training set to learn the characteristics of the different classes and builds up a model that tries to capture these characteristics. After the training phase, the classifier can be seen as a function $C : D \mapsto Y$. That is, the classifier takes an instance $i \in D$ and assigns a class label $l \in Y$ to it.
A well-known example is the classification of incoming emails into either spam or non-spam. In this scenario, a classifier gets a set of mails that are labelled as spam or as non-spam. With this set, the classifier tries to learn the characteristics of both classes. For example, spam mails often contain many words written in capital letters, while non-spam mails do not. With the knowledge learned in the training phase, the classifier can then be used to identify spam mails, i.e., by assigning the class label "spam" to them.
The most common parameter to evaluate a classifier is the *accuracy* which is the percentage of correctly assigned class labels. To determine the accuracy, the classifier is given a labelled set of instances. For each instance, the label assigned by the classifier is then compared to the actual label. There exist many more evaluation parameters that we do not cover here. An overview as well as selection strategies is given in [LZWT14].

A major problem of supervised classification is *overfitting*. This phenomenon describes that the classifier learns the instances of the training set by heart, but does not learn generalisable characteristics and thus has very low accuracy on new, unknown instances. In the mail example, this would be the case if the classifier learns all given spam mails in detail. That way, it would correctly identify these mails as spam, but minor variations in the mail would cause it to not identify them correctly. For that reason, evaluation of classifiers (e.g., determining the accuracy) should not be done on the set used to train the classifier, as it can greatly overestimate the quality of the classifier. Instead, a common approach is to split the labelled instances into a training set and a test set, where the latter is used for evaluation.

2.2.2. The Decision Tree Classifier

In this chapter, we introduce a specific classifier that we will use in a later chapter. There are many well-known approaches of classifiers. As stated by Wolpert et al. [WM97] in their "No free lunch theorem", neither of these approaches can claim to be the best classifier; each of them has different strengths and weaknesses, also depending on the specific application. For our gas network classification task, we decided that the *Decision Tree Classifier* seems to be well-suited for multiple reasons. It can be built manually, one can prioritise important attributes very well, and it can be adapted easily. Furthermore, it results in decision rules

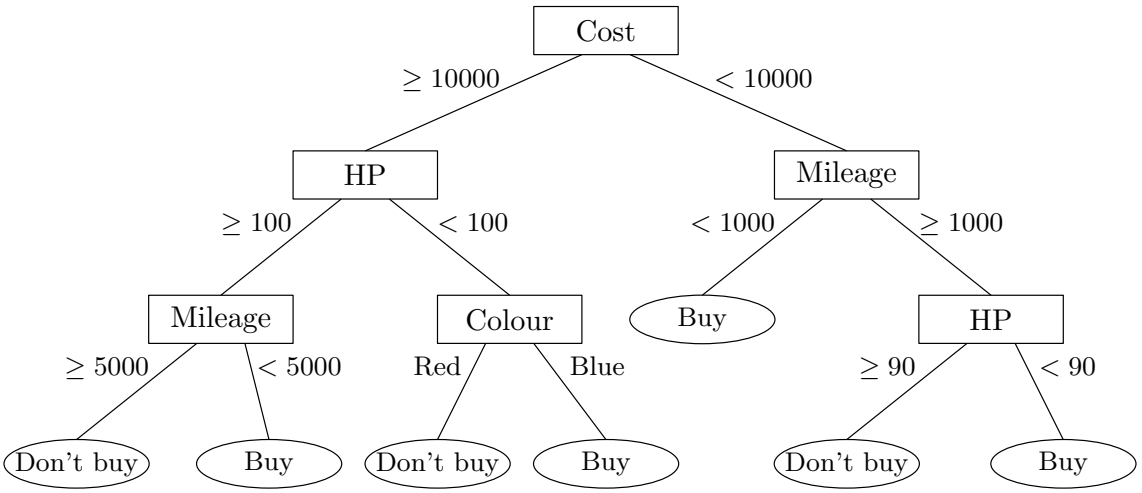


Figure 2.1.: Example decision tree. The attributes are the cost in euro, the mileage in kilometres, the horsepower (HP) and the colour.

that are easy to understand. Therefore, the classification results can be retraced, making it easier to find possible mistakes.

The idea of decision trees is to successively choose attributes and split the training set instances in multiple subsets, depending on the value of the chosen attribute in each instance. An attribute can be split into intervals if it is numeric. For example, a split based on the attribute *length* and a threshold x can be "$length < x$ or $length \geq x$". If the attribute is categorical, like the attribute *colour*, it is split based on the possible values, for example "is the colour *green*, *red* or *blue*". On each branch of the tree, the next attribute to split can be chosen independently, and not all attributes have to be used. The recursion on a branch ends if some stopping criterion is fulfilled, e.g., all remaining instances (or at least the majority) within the branch belong to the same class. In this case, a leaf is created with the corresponding class label. An example decision tree is shown in Figure 2.1. The task in this example is to answer the question if one should buy a specific car with the numeric attributes *cost, mileage* and *horsepower* and the categorical attribute *colour*. The class labels are *Buy* and *Don't Buy*. We see that the order of split attributes can be chosen freely for each branch and that branches do not have to use all or even the same number of attributes.

A constructed decision tree can be seen as a set of rules, where every path from the root to a leaf is the conjunction of $IF - THEN$ rules. To classify an instance, one starts at the root of the tree and checks at each inner node the value of the split attribute to follow the corresponding branch. The leaf then provides the predicted class label. We go through a small example with the decision tree shown in Figure 2.1. The car to be classified is blue, costs €15000, its horsepower is 120 and it has a mileage of 3000 km. At the first node (which is the root) we compare the cost of €15000 to the split threshold of €10000 and see that they are greater, thus we follow the left branch. The same is true for the second inner node which checks the horsepower. At the last inner node, the mileage of 3000 km of the car is smaller than the threshold of 5000 km, thus we follow the right branch. We have then reached the leaf of the branch with the label *Buy*. That label is now assigned to the car in question.

One advantage of decision trees is that one can sort the attributes by their importance. Different ways exist to identify the most important or useful attributes, like the information gain or the Gini index [Tan20]. Probably the biggest disadvantage of decision trees is the

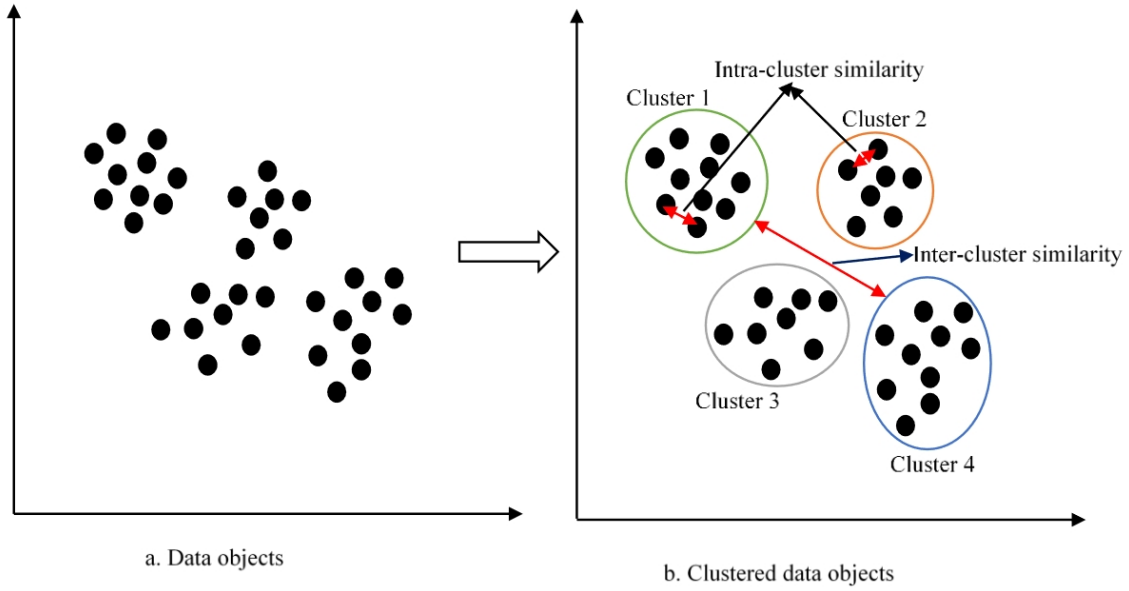


Figure 2.2.: Example clustering of two-dimensional objects [ESA+21].

previously mentioned overfitting. To prevent this, the tree often gets pruned, either while constructing it or after it is built [Min89].

2.2.3. Relation to Clustering

Clustering can be seen as the unsupervised equivalent to classification. In this problem, the set of classes Y is not known upfront, hence there is also no labelled training set. Thus, the task is not only to find a function $C : D \mapsto Y$, but also to find the class set Y. Clustering algorithms try to find common features between instances and cluster them in a way that instances within the same cluster are very similar to each other, while instances in different clusters significantly differ from each other. An example clustering of two-dimensional objects (taken from [ESA+21]) is shown in Figure 2.2.
As for classification, there are many different approaches of clustering algorithms (see [Agg18]) that we do not discuss in this work.

2.3. Statistical Parameters

In this section, we explain several widely used parameters from descriptive statistics: the minimum and maximum, the arithmetic mean and the median, and the standard deviation. All of these parameters are used to summarize specific characteristics of a variable in a data set D in one single scalar.
For the notations in the following explanations, assume a variable x, a set $M \subseteq D$ and a set of $|M|$ values $X = (x_1, x_2, ..., x_{|M|})$. Each x_i is the value of x in the i-th element of M. If $M = D$, i.e., all elements of the data set are included in the calculation of a parameter, we call M the *population*. In contrast, if only a subset of all elements is used (i.e., $M \subset D$), we call M a *sample*.
In graphs, an example of a variable x is the attribute *length* of an edge. The set $L = (l_1, l_2, ..., l_{|E|})$ is then the set of edge lengths in the set $M = E$.
In the following, we assume that the variable x is numeric. Thus, an ordering of its values x_i exists and basic math operations like addition and division are defined for two elements x_i, x_j.

The parameters *minimum* and *maximum* denote the smallest and the largest value that x takes in M:

$$x_{min} = \min_{x_i \in X} x_i, \quad x_{max} = \max_{x_i \in X} x_i. \tag{2.2}$$

The *arithmetic mean* is one way of describing the average value of x in M. It is computed as the sum of all values divided by the number of values:

$$\overline{x} = \frac{1}{|M|} \sum_{i=1}^{|M|} x_i. \tag{2.3}$$

If the values x_i are weighted with a scalar w_i, we can also compute the *weighted arithmetic mean*:

$$\overline{x}_w = \frac{\sum_{i=1}^{|M|} w_i \cdot x_i}{\sum_{i=1}^{|M|} w_i}. \tag{2.4}$$

Another way of describing the average value of x in M is the *median* which is "the value in the middle" when looking at the values x_i in ascending order. Formally, the median is defined as

$$x_{mdn} = \begin{cases} x_{\frac{|M|+1}{2}}, & \text{if } |M| \ mod \ 2 = 1 \\ \frac{1}{2}(x_{\frac{|M|}{2}} + x_{\frac{|M|}{2}+1}), & \text{otherwise.} \end{cases} \tag{2.5}$$

While both the arithmetic mean and the median describe the average value of x, they vary to a certain extent. The most notable difference is that the median is more robust against *outliers*. These are values in X that are significantly higher or lower than the majority of all values. Outliers are often considered insufficiently representative of the data, so it is not desirable for them to influence the parameters too much. As it is often done in literature, when talking about the *average* (or using the abbreviation *avr*) of a variable, we mean the arithmetic mean.

The last parameter is the *standard deviation* that describes the average difference between the values x_i and the arithmetic mean $\overline{x}$. A higher standard deviation value indicates that the values x_i are widely spread around $\overline{x}$, whereas a lower value indicates that most values are close to $\overline{x}$. Depending on M being the population or a sample, there are two definitions of the standard deviation:

$$stdev_{population} = \sqrt{\frac{1}{|M|} \sum_{i=1}^{|M|} (x_i - \overline{x})^2}, \tag{2.6}$$

$$stdev_{sample} = \sqrt{\frac{1}{|M|-1} \sum_{i=1}^{|M|} (x_i - \overline{x})^2}. \tag{2.7}$$

Since we always use $stdev_{population}$ in our analyses, from now on we will only use the notation $stdev$, by which we mean $stdev_{population}$.

3. Gas Network Model and Problem Statement

In this chapter, we explain how gas networks are modelled and how we transform this model into a graph representation. With this model and the previous chapter in mind, we formally define the problem we solve in this work. We also give a brief description of the expected different gas network classes.

3.1. Gas Networks and Properties of our Data Set

Gas networks are connected *gas pipelines* that each transfer gas from one *intersection point* to another. In addition to these two components, gas networks also have *regulators* which are responsible for down-regulating the pressure. The pressure is an important property of gas networks and refers to the pressure at which the gas is transferred through a pipeline. During the gas transfer, the pressure naturally decreases slowly from the start point of the pipeline to its end point. At several points the pressure has to be down-regulated to a desired value that is often significantly lower than the current value. This is accomplished by the regulators. These constructs are placed within an intersection point and have an input, where the gas arrives with the old pressure value, and an output, where the gas leaves with the new pressure value.
We model a gas network as a connected graph, where the intersection points are modelled as nodes and the gas pipelines are modelled as edges. In the given data sets, regulators are also modelled as edges. Each regulator is associated with a start node, an end node and the newly regulated pressure. We refer to the pipeline edges as the set E_p and to the regulator edges as the set E_r. Together with the node set V, this results in our graph $G = (V, E_p \cup E_r)$. We refer to the number of nodes as n and to the number of pipeline and regulator edges as m_p and m_r, respectively.
While gas flows actually have a direction in reality, we choose to model the network as an undirected graph since the orientation of the edges in our data set have been set arbitrarily. Still in some cases, we may talk about flow direction. In most cases the direction is indicated by the pressure decline along the edge. This is particularly true for regulator edges for which start node and end node are defined.
Each node is associated with two attributes. The first one is the pressure (given in bar) that the gas flow has at this node, the second one denotes if the node is considered a *source* of the network. That is if the node is either a source node of the whole network or the end node of a regulator. For $u \in V$, we refer to these attributes as $p(u)$ and $s(u)$.

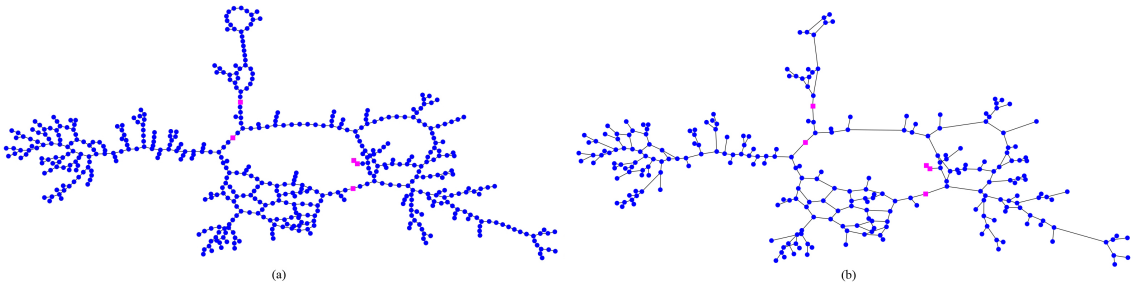


Figure 3.1.: Example gas network. (a) Shows the original graph, (b) shows the contracted graph.

Each pipeline, and therefore each $e \in E_p$, has two attributes: its inner diameter, given in millimetres, and its length, given in meters. We refer to these values as $i(e)$ and $l(e)$. In our data set, no pressure was provided for the pipelines. Since we are often interested in that parameter, we approximate it as the average of the pressures of its associated nodes. Thus, we denote the pressure of a pipeline edge $e = \{u, v\}$ as $p(e)$ and compute it as $p(e) = \frac{p(u)+p(v)}{2}$.
Each regulator $r = (u, v)$ has only one attribute, that is the pressure $p(r)$ to which the gas flow is regulated. The regulation is performed between u and v, thus $p(v) = p(r)$.
An example gas network (or rather its graph representation) is shown in Figure 3.1 (a). The pink squares illustrate nodes that are sources of the network, all other nodes are shown as blue cycles. Edges between nodes are shown as lines. Note that our given gas networks do not provide coordinates. Thus, the visualisations show how the nodes are connected among each other, but they do not reveal how the nodes are located in relation to each other.

Before we proceed, we state some things concerning our view on regulators. Although they are modelled as edges between two nodes in the data, they do not really fulfil the purpose and characteristics of an edge: that is, connecting two points at different locations. This purpose is fulfilled by the pipeline edges, that also have typical edge attributes like the length. In contrast, regulators are (as stated before) a construct belonging to one node, altering the pressure between the input and the output of the node. Hence, from now on, when talking about the edges of G, we implicitly mean *only* the pipeline edges E_p if not explicitly stated otherwise. Furthermore, when talking about a regulator, we mean its output node. However, in our visualisations a regulator is still shown as an edge between two nodes (see Figure 3.1), with the end node illustrated as a pink square.

Furthermore, the given gas networks in our data set often have nodes with degree two which exist only for modelling reasons but do not hold any information about the represented gas network. These artificial nodes and edges make some analyses blurry. For example, the average length of edges significantly decreases when an edge is divided into two edges by an artificial node. Therefore, we contract these nodes with degree two. The only exception to this are regulators since they carry important information. Contracting a node v with incident edges $\{u, v\}$ and $\{v, w\}$ means that v and its incident edges are removed and replaced by an edge $\{u, w\}$. The length of the new edge is the sum of the lengths of the removed edges. The inner diameter of both removed edges should be the same (since they represented a single pipeline), but due to some errors in the data, for very few edges this was not the case; for these edges, we take the average of the inner diameters of both removed edges.
These contractions are made repeatedly until no nodes with degree two exist in G. The resulting contracted graph of the previous example is shown in Figure 3.1 (b). All further examinations are performed on this contracted graph.

Another detail worth mentioning is that in some gas network instances a very small part of the nodes has erroneous values for the pressure attribute, namely a value of zero. We do not include these nodes in computations that use the pressure attribute.

3.2. Problem Statement

With the previous chapters in mind, we formally define our gas network classification task and provide some further details.
We are given an *expected* set of classes $Y = \{$*Inner City, Old Town, Residential Area, Rural Area, Industrial Area*$\}$. These classes are gas network types we expect to exist in reality, meaning that each of them has its own characteristics so that we can distinguish between them. In the best case, the characteristics of all networks of the same type are very similar, but very dissimilar across the different types. It is important to note that this set is not fixed and we do not know beforehand if we can confirm the existence of each class. It is very well possible that some classes have similar characteristics so that they should be only one class, and it is also possible that we find even more classes. On the other hand, we cannot rule out to discover that nearly all networks are so different from each other that we cannot categorize them at all.
In addition to the set of classes, we are also given a data set of gas networks, each of them labelled with exactly one class label. It is worth mentioning that all observations and results in this work depend on this data set. Given more data, these results have to be reviewed and possibly refined.
With this said, the task can be divided into two parts, although they are of course not strictly separated. The first part is to either confirm our expected class set Y or find another meaningful class set. The second part is to build one or more classifiers that assign these class labels to new gas networks. For both tasks, we analyse the given labelled gas networks to identify graph parameters that offer valuable information about differences and common features between the classes.
This is also why we introduced the problem of clustering: while we expect our task to be a classification problem with known labels, it is possible that these classes in fact do not exist and we have to find new classes based on the characteristics of the networks. If the latter case is true, our task transforms into a clustering problem.

3.3. Descriptions and Expectations of the Network Types

In this section, we describe the gas network types introduced in the previous section. We first explain the areas which the networks in each class originate from. Then, we present some assumptions and expectations we have regarding the different classes and their distinguishability.

3.3.1. Description

Our first class label is *Inner City*. With this we refer to the city centre of large cities.
With the class *Old Town* we refer to the historic centre of cities. These tend to exist more in larger cities rather than in small towns.
Next is the class *Residential Area*. These are areas like small towns and large villages.
The next class is *Rural Area*. These are larger regions in rural areas. They often include a main supply line from which several small villages branch off at certain intervals.
The last class is *Industrial Area*. These are areas mainly used for industrial development and are usually located at the edge of cities or in rural areas.

An example gas network of each class (in the above order) is shown in Figures 3.2, 3.3, 3.4, 3.5 and 3.6.

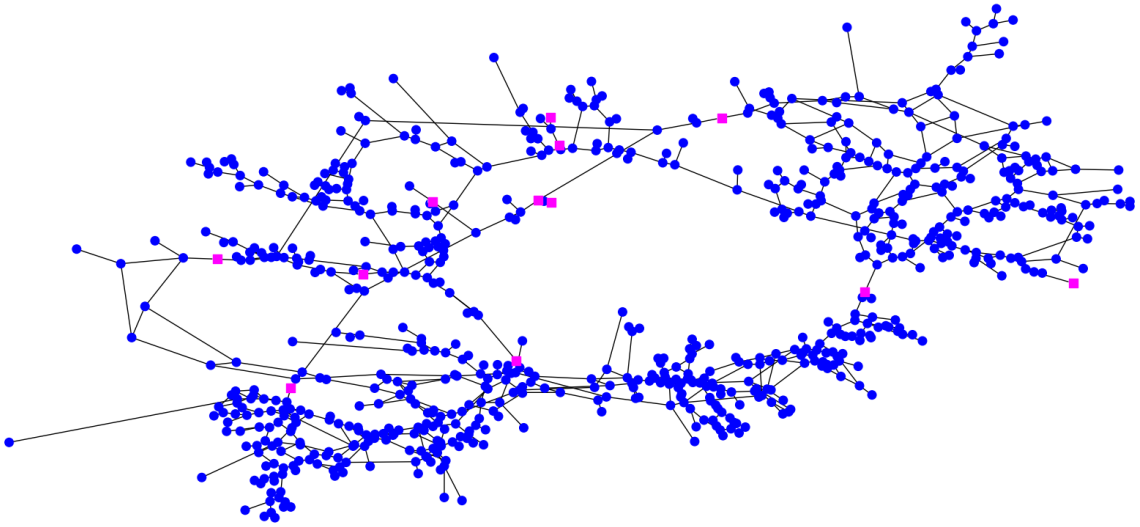


Figure 3.2.: Example gas network with the label *Inner City*.

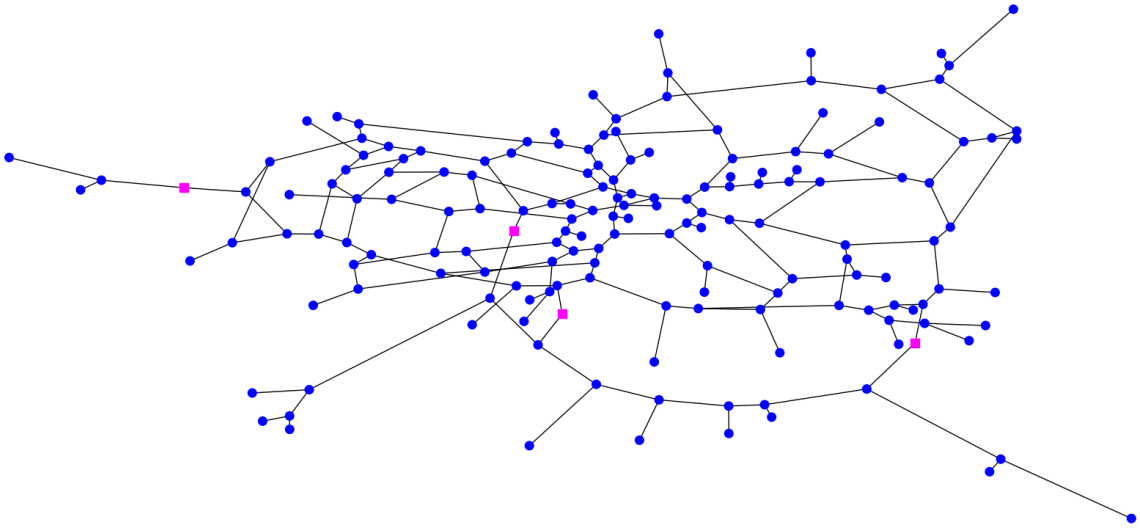


Figure 3.3.: Example gas network with the label *Old Town*.

3.3.2. Assumptions and Expectations

With these class descriptions in mind, we expect that the largest differences in the topology of networks are found between rural area networks and inner city networks. This is due to the main characteristics we assume in these areas. Usually, inner cities are very densely populated, that means there are a lot of buildings in relation to the size of the area. In rural areas the distances between buildings are usually a lot higher. Also, since rural areas often cover several villages and the area between them, we expect their geographic extent to be larger than the extent of inner cities. Overall, inner cities tend to be more compact than rural areas. Furthermore, inner cities tend to be more meshed than rural areas in terms of street connections and crossings. We expect these characteristics to be represented in the topology of their gas networks and therefore expect a fairly high degree of distinguishability between networks of these classes.

Thus, we expect these two classes to be the extremes with regard to the mentioned characteristics. For old towns and residential areas we expect them to be "between" these extremes. Residential areas for example are usually less densely populated than inner cities,

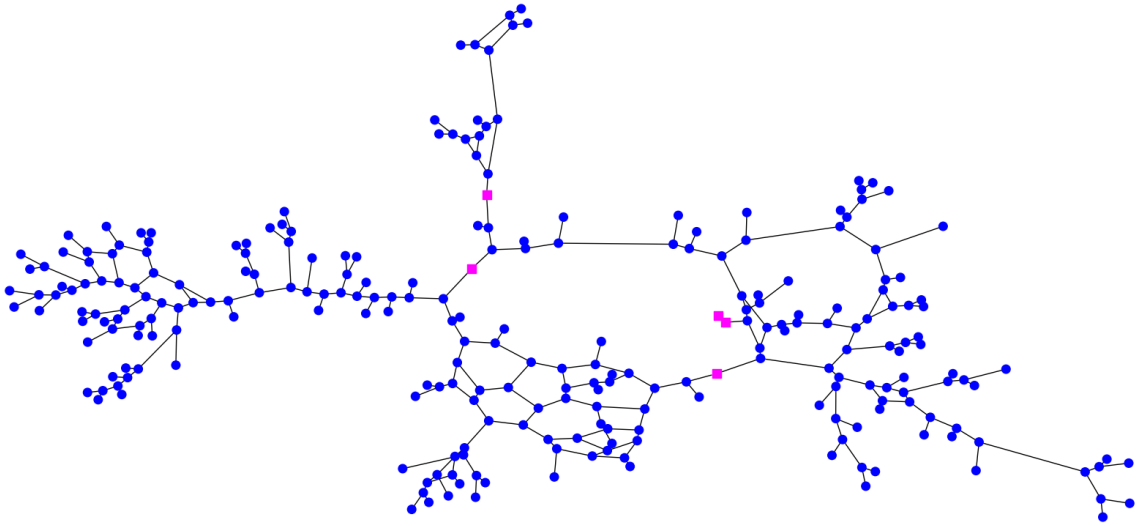


Figure 3.4.: Example gas network with the label *Residential Area*.

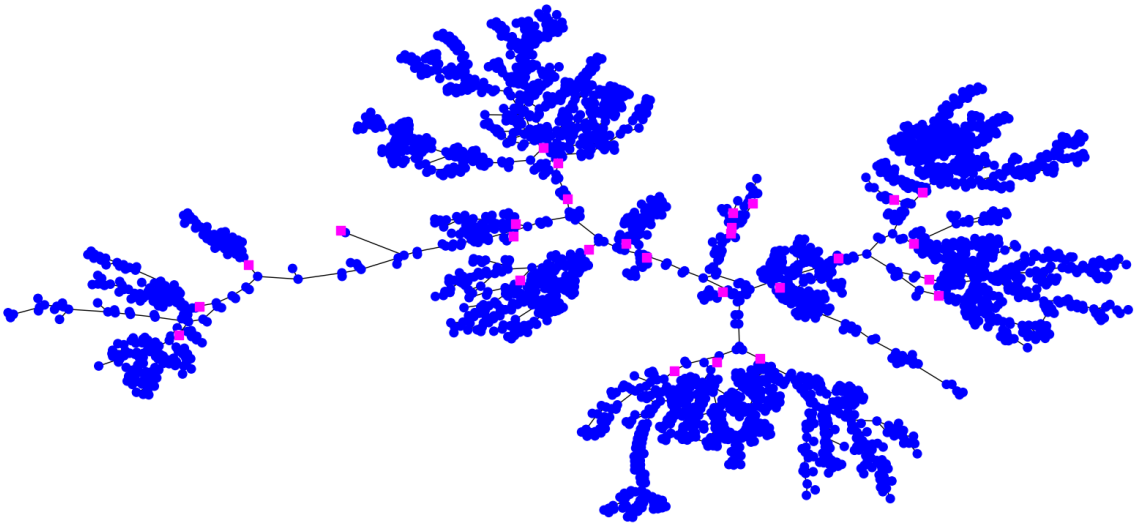


Figure 3.5.: Example gas network with the label *Rural Area*.

but still more than rural areas. As for the differences between residential areas and old towns we do not have specific assumptions.

The characteristics of industrial areas are quite unknown to us upfront, therefore we cannot provide any well-grounded assumptions on these networks. However, we expect that in some way industrial areas are structured differently than the other classes, since they fulfil a different purpose.

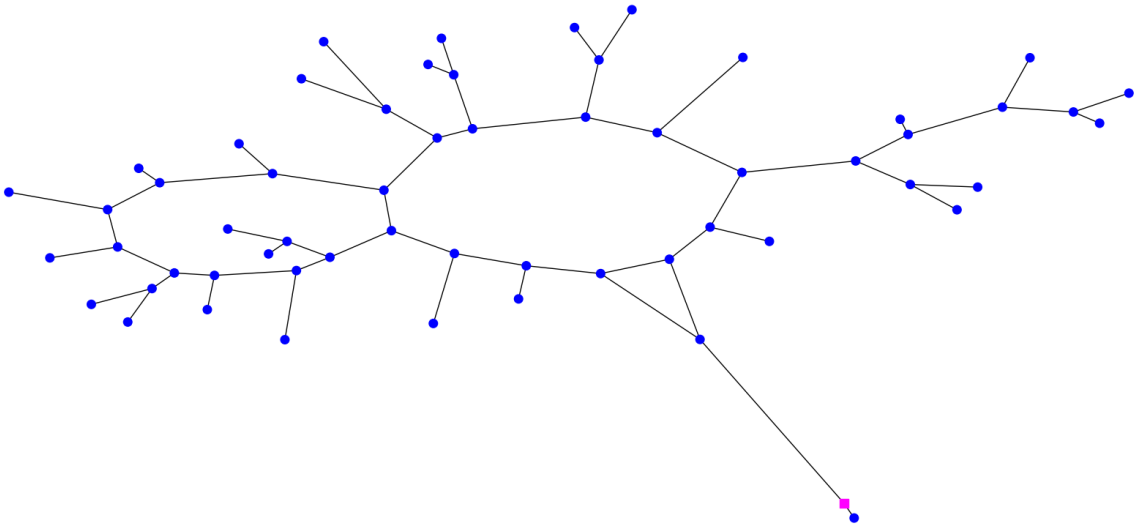


Figure 3.6.: Example gas network with the label *Industrial Area*.

4. Examined Parameters

In this chapter, we present the parameters we selected to analyse and describe the characteristics of the gas network instances and classes. For each parameter, we give an explanation what the parameter is in terms of graph theory and what it says about a graph in general. Then, we state what the parameter says in the context of gas networks. For some parameters, we also give some insight in what we expect from it. This includes an explanation of why we think that this parameter can be meaningful to describe the networks, as well as an estimation of the parameter values in the different gas network classes. For these estimates, we often refer to the expected extremes mentioned in the previous chapter, i.e., the inner cities and the rural areas.
The parameters are grouped in five sections. We provide a tabular overview of all parameters in Table 4.1. It contains a brief description of each parameter as well as references to the corresponding subsection. Note that each examined parameter P refers to the graph G, thus for better readability we only write P instead of $P(G)$.
In this and the following chapters, we will use the terms *gas network* and *gas network graph* synonymously, i.e., when talking about parameters of a gas network, we mean the parameters of its graph representation and vice versa. Similarly, we use the terms pipeline and edge synonymously, depending on the context. Also remember that when talking about the edges of a network, we implicitly mean only the pipeline edges E_p, unless stated otherwise.

Before we start explaining the selected parameters, we briefly explain why we choose to not use the number of nodes and edges explicitly to describe the characteristics of the gas networks. While it is somewhat intuitive to use the number of nodes as the "size" of a graph, this is not necessarily true for gas networks. The size of a gas network would rather be seen as the geometric area it covers, and the number of nodes does not tell too much about that area. Hence, we do use the number of nodes and edges for some parameters, but do not use them as own parameters.

4.1. Length, Geographic Extent and Inner Diameter

4.1.1. Length of Pipelines

Each edge has an attribute *length* that indicates how many meters the pipeline is long. We analyse four statistical parameters of the values of this attribute: the minimum and maximum length of a pipeline in the graph, the arithmetic mean and the median of all

Table 4.1.: Overview of all parameters examined in this chapter.

Parameter	Notation	Description	Ref.
Length	Len_{min}, Len_{max}, Len_{avr}, Len_{mdn}	Statistics about the lengths of edges	4.1.1
Average path length	PLavr, PLavr_{n}	Average shortest path distance between all pairs of nodes	4.1.2
Diameter	Diam, Diam_{n}	Maximum eccentricity among all nodes	4.1.2
Radius	Rad, Rad_{n}	Minimum eccentricity among all nodes	4.1.2
Centrality	Cent, Cent_{n}	Average distance from all nodes to the most central node	4.1.2
Inner diameter	ID_{min}, ID_{max}, ID_{avr}, ID_{mdn}	Statistics about the inner diameter of edges	4.1.3
Node degree	Deg_{min}, Deg_{max}, Deg_{avr}	Minimum, maximum and average node degree	4.2.1
Clustering coefficient	CC	Average clustering coefficient of all nodes	4.2.1
Minimum and maximum pressure	Pr_{min}, Pr_{max}	Average pressure of the pipelines	4.3.2
Average pressure	Pr_{avr}	Average pressure of the pipelines, weighted with their length	4.3.2
Number of pressure stages	NumStages	Number of different pressure stages	4.3.2
Number of pressure areas	NumAreas	Total number of pressure areas	4.3.2
Regulators	NumRegulators	Total number of regulators	4.3.3
Source areas	NumSources	Number of source areas	4.3.3
Sink areas	NumSinks	Number of sink areas	4.3.3
Area supplies: regulators	$\text{NumSuppliesTotal}_{\text{max}}$, $\text{NumSuppliesTotal}_{\text{avr}}$	Number of regulators that supply an area	4.3.3
Area supplies: areas	$\text{NumSuppliesAreas}_{\text{max}}$, $\text{NumSuppliesAreas}_{\text{avr}}$	Number of pressure areas that supply an area	4.3.3
Area supplies: stages	$\text{NumSuppliesStages}_{\text{max}}$, $\text{NumSuppliesStages}_{\text{avr}}$	Number of different pressure stages that supply an area	4.3.3
Skipped pressure stages	$\text{SkippedStages}_{\text{max}}$, $\text{SkippedStages}_{\text{avr}}$	Maximum and average number of skipped pressure stages	4.3.3
Low pressure graph parameters	$\text{Len}_{\text{low}_{\text{max}}}$, $\text{Len}_{\text{low}_{\text{avr}}}$, $\text{ID}_{\text{low}_{\text{max}}}$, $\text{ID}_{\text{low}_{\text{avr}}}$, $\text{Deg}_{\text{low}_{\text{max}}}$, $\text{Deg}_{\text{low}_{\text{avr}}}$	Parameters on the lower pressure stages	4.3.5
Relative size of the 2-core	SizeTwocore	Proportion of nodes that belong to the 2-core	4.4.2
Origin nodes on core paths	$\text{NumOrigins}_{\text{min}}$, $\text{NumOrigins}_{\text{max}}$, $\text{NumOrigins}_{\text{avr}}$, $\text{NumOrigins}_{\text{stdev}}$	Statistics about the number of origin nodes on core paths	4.4.2
Distances between origin nodes	$\text{DistancesOrigins}_{\text{min}}$, $\text{DistancesOrigins}_{\text{max}}$, $\text{DistancesOrigins}_{\text{avr}}$, $\text{DistancesOrigins}_{\text{stdev}}$	Statistics about the distances between origin nodes	4.4.2
2-core nodes with high degree	RatioHigherDegree	Proportion of nodes in 2-core that have degree ≥ 3	4.4.2
Origin nodes with high degree	OriginsHigherDegree	Number of origin nodes with degree ≥ 3	4.4.2
Multiple origins	OriginsMultiple	Number of origin nodes that are origin of more than one tree	4.4.2
Size of trees	$\text{SizeTrees}_{\text{min}}$, $\text{SizeTrees}_{\text{max}}$, $\text{SizeTrees}_{\text{avr}}$, $\text{SizeTrees}_{\text{stdev}}$	Statistics about the size (number of nodes) of the attached trees	4.4.3
Depth of trees	$\text{DepthTrees}_{\text{min}}$, $\text{DepthTrees}_{\text{max}}$, $\text{DepthTrees}_{\text{avr}}$, $\text{DepthTrees}_{\text{stdev}}$	Statistics about the depth of the attached trees	4.4.3
Leaves of trees	$\text{LeavesTrees}_{\text{min}}$, $\text{LeavesTrees}_{\text{max}}$, $\text{LeavesTrees}_{\text{avr}}$, $\text{LeavesTrees}_{\text{stdev}}$	Statistics about the number of leaves of the attached trees	4.4.3
Treewidth	TW	Treewidth of the graph	4.5
Planarity	IsPlanar	True if and only if the graph is planar	4.5

edge lengths. We denote these as $Len_{min}(G)$, $Len_{max}(G)$, $Len_{avr}(G)$ and $Len_{mdn}(G)$, respectively.
We expect to observe higher maximum and average values of pipeline lengths in rural areas, since there the distances between pipeline intersections are longer than in more densely populated areas. For the minimum we do expect quite similar values across all network types.

4.1.2. Path Lengths, Network Extent and Centrality

This subsection covers parameters that describe the geographic extent and the shape of a graph. For this use, we first define the distance $d(u, v)$ between two nodes u, v, also referred to as path length between u and v. That distance is the minimum path length among all possible paths from u to v, using the edge length as metric. For a node v, we call the maximum distance to any other node the eccentricity $ecc(v)$ of v. With these definitions, we introduce three parameters: the average path length $APL(G)$ (Equation 4.1), the diameter

$Diam(G)$ (Equation 4.2) and the radius $Rad(G)$ (Equation 4.3) of a graph G, which are defined as follows:

$$\text{APL(G)} = \frac{1}{n \cdot (n-1)} \sum_{u,v \in V, u \neq v} d(u,v), \tag{4.1}$$

$$\text{Diam(G)} = \max_{v \in V} ecc(v), \tag{4.2}$$

$$\text{Rad(G)} = \min_{v \in V} ecc(v). \tag{4.3}$$

The average path length is the arithmetic mean of the distances between each pair of nodes in G. The graph diameter is the maximum distance of two nodes in G and describes the extent of G. Since we expect rural areas to cover larger areas, we expect this value to be higher in these networks. The graph radius is the minimum eccentricity among all nodes in G. That means that from the most central point in G (with regard to the eccentricity), the distance to all other nodes is less than or equal to $Rad(G)$.

With the next parameter, we try to further describe the shape of the graph by determining its centroid c and the average distance of all nodes to c that we refer to as the centrality $Cent(G)$ of the graph. As the centroid c, we define the node for which the average distance to all other nodes is minimal, which leads to the following definition of $Cent(G)$:

$$\text{Cent(G)} = \min_{c \in V} \left\{ \frac{1}{n-1} \sum_{v \in V, v \neq c} d(c,v) \right\}. \tag{4.4}$$

Note that in contrast to the definition of the radius, the centricity uses the average path lengths to all other nodes instead of the maximum. It is therefore another way of defining the centre of G.
Due to the expected shape of rural areas, which is a long backbone with several detouring branches, we expect both the centrality and the radius to be much higher than in inner cities that are often build around a central core.

In addition to each parameter P described in this subsection, we also introduce a parameter P_n that is normalised by the number of nodes n of G, i.e., $P_n(G) = \frac{P(G)}{n}$. These normalised parameters allow for better comparison of networks with different numbers of nodes. Furthermore, they give some approximations of the node density. Take for example a small rural area and a large inner city: while both could have a similar diameter, we expect the density of nodes to be much higher in the inner city and this would be apparent as a smaller diameter per node value.

4.1.3. Inner Diameter of Pipelines

The other attribute associated with each edge is the inner diameter. We analyse the same statistical parameters as for the length, so the minimum and maximum value as well as the arithmetic mean and the median of all edges. We denote these as $ID_{min}(G)$, $ID_{max}(G)$, $ID_{avr}(G)$ and $ID_{mdn}(G)$, respectively.
The inner diameter is an indicator on the upper bound of gas that can be transferred through a pipeline in a fixed amount of time. However, it is only one part of the physical parameters that bound the gas flow, and it does not contain much information about the actual amount of gas transferred there. Nonetheless, we expect its value to be higher in networks like inner cities that are more densely populated.

4.2. Meshedness and Connectivity

In this section, we cover several parameters that describe how meshed a graph is. These parameters are heavily based on the number of nodes and edges in the graph as well as on

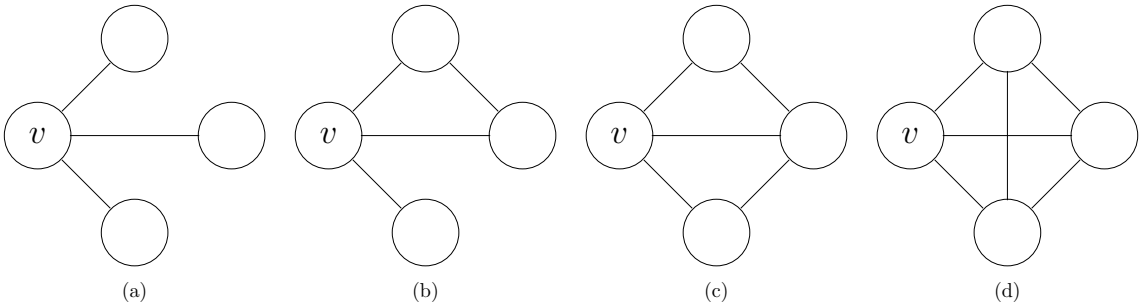


Figure 4.1.: Clustering coefficient of node v. (a) $cc(v) = 0$, (b) $cc(v) = \frac{1}{3}$, (c) $cc(v) = \frac{2}{3}$, (d) $cc(v) = 1$.

the node degrees. We first explain all parameters and then give an overall expectation of the resulting values in our gas network classes.

4.2.1. Parameters

The first three parameters are the minimum and maximum node degree $Deg_{min}(G)$ and $Deg_{max}(G)$ and the average node degree $Deg_{avr}(G)$, which are defined as follows:

$$\mathrm{Deg_{min}(G)} = \min_{v \in V} d(v), \tag{4.5}$$

$$\mathrm{Deg_{max}(G)} = \max_{v \in V} d(v), \tag{4.6}$$

$$\mathrm{Deg_{avr}(G)} = \frac{\sum_{v \in V} d(v)}{n}. \tag{4.7}$$

The average node degree is directly related to another common parameter, namely the edge node ratio. This parameter is computed as the number of edges divided by the number of nodes. Since this equals exactly half of the average node degree, we do not consider it as an additional parameter in our further analyses.

Next is the average clustering coefficient $CC(G)$ of a graph G. To define it, we first need to define the clustering coefficient $cc(v)$ of a node $v \in V$:

$$\mathrm{cc(v)} = \frac{2e_v}{d(v)(d(v)-1)}, \tag{4.8}$$

where e_v is the number of edges between the nodes incident to v. Thus, the clustering coefficient for a node v is the ratio of the actual number of edges between its neighbours and the maximum possible number of edges between them. In graph context, this value says how close the neighbours of v are to being a clique. Of course, $cc(v) \in [0, 1]$. An example is shown in Figure 4.1: node v has three neighbours, so at most three edges can exist between these neighbours. The actual number of edges e_v increases from 0 in (a) to 3 in (d), and therefore the clustering coefficient of v increases from 0 in (a) to 1 in (d).
We can now formally define the parameter $CC(G)$ as the arithmetic mean of all node clustering coefficients:

$$\mathrm{CC(G)} = \frac{1}{n} \sum_{v \in V} \mathrm{cc(v)}. \tag{4.9}$$

4.2.2. Expectations

All of the mentioned parameters, except of Deg_{min}, have higher values in more tightly meshed graphs. Our expectation in general is that inner cities and residential area networks

are more meshed than rural areas and industrial areas. Thus, in the former we expect higher values for most parameters.
In terms of node degrees, we expect the average node degree to be the most meaningful. For Deg_{min} we expect the value 1 for each network since it is very likely that each network has at least one leaf. For the maximum degree, inner cities could show higher values than rural areas, but we can also imagine that the node degree is bounded to 4 for all gas networks.

4.3. Pressure

In this section, we introduce parameters related to the pressure of the gas network nodes. Since this attribute is directly associated to the context of gas networks, our explanations are related to these networks rather than to graphs in general.
As stated in Chapter 2, the pressure is decreased in two ways. First, it naturally decreases slowly while the gas is transferred through the network. Second, and more relevant, the pressure can be significantly decreased by regulators. We also stated that the pressure attribute is associated with the network nodes, but we can approximate and therefore talk about the pressure on an edge.
In the following, we will introduce our concepts of pressure stages and pressure areas. With these definitions, we then explain the parameters we selected to describe the characteristics of networks with regard to the pressure. At last, we give again some expectations concerning the parameter values in the different network types.

4.3.1. Introduction into Pressure Stages and Pressure Areas

Among all nodes, lots of different pressure values exist. When sorting these values in decreasing order, we observe two major categories of steps between two adjacent values. Most steps are very small (e.g., a decrease from 2.115 to 2.114) and derive from the decrease in pressure during the gas transfer. Then there are some larger steps (e.g., a decrease from 8.0 to 5.0) that derive from a regulator decreasing the pressure. With the concept of pressure stages, we try to divide the pressure values into intervals corresponding to these large steps. Each interval is then considered as a pressure stage and represented by its upper bound. The algorithm that computes these intervals is shown in Algorithm 4.1. In the first step (lines 1-4) we extract the pressure values from the nodes and sort them in decreasing order. The list of pressure stages (represented by upper bounds) is initialized as an empty list (line 5). The variable *previous pressure* always holds the previously seen pressure value. To ensure that the first value indicates a large step, this variable is initialized with infinity (line 6). Now the main loop begins in which we iterate over all pressure values in decreasing order (line 7). For each value, we check if we identify a large step compared to the previous value. We define such a large step as a value decrease by the factor 0.8 or smaller (line 8). If this condition is true, we have found a new pressure stage and add the current value as an upper bound to the list of pressure stages (line 9). If the condition is not true, the current value does not belong to a new pressure stage. In both cases, we update the previous seen pressure to the current one and proceed with the next value (line 10). This is done until all values have been checked, then the list of pressure stages is returned (line 12).

The pressure stage of each node $v \in V$ is then the interval that $p(v)$ fits into, i.e., $p(v)$ is less or equal than its upper bound, but greater than the following upper bound.
In Figure 4.2, the resulting pressure stages of an example gas network are visualised. All nodes of the same colour belong to the same pressure stage, pink square nodes are sources of the network (remember that this includes the end nodes of regulators). In this and following visualisations of pressure stages, the ordering from highest to lowest pressure stage is *red*, *orange*, *yellow*, *blue*, *green*, *purple*, *brown*, *grey*. Remember that due to errors some nodes have a pressure value of zero; these are shown as black nodes and we do not

Algorithm 4.1: COMPUTING THE PRESSURE STAGES OF A NETWORK G

Input: Graph $G = (V, E)$, pressure $\mathsf{p} : V \to \mathbb{R}$
Output: List of pressure stages, represented by their upper bound

```
1  pressures ← ∅
2  forall v ∈ V do
3      pressures.APPEND(p(v))
4  sortedPressures ← SORTDECREASING(pressures)
5  pressureStages ← ∅
6  previousPressure ← ∞
7  forall pressure ∈ sortedPressures do
8      if pressure ≤ previousPressure · 0.8 then
9          pressureStages.APPEND(pressure)
10     previousPressure ← pressure
11
12 return pressureStages
```

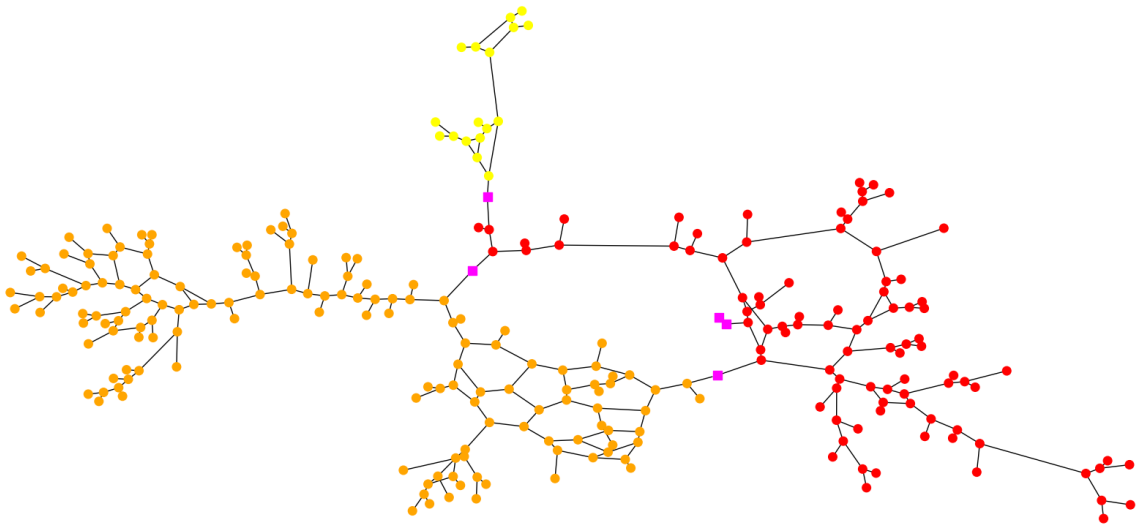


Figure 4.2.: Pressure stages of a gas network. All nodes of the same colour belong to the same pressure stage; pink nodes represent sources of the network.

consider them as a pressure stage. Also note that the colours are not related to concrete pressure values and therefore cannot be compared among networks. They only show the ordering of pressure stages.
Using the example in Figure 4.2, let us briefly recap the meaning of these pressure stages. All nodes of the same colour nearly have the same pressure, only differing slightly due to the decrease in pressure while the gas is transferred through the network. Between different stages, large differences in pressure exist. The different stages are strictly separated from each other, with the only exception being connections by regulators. This follows from the way we constructed the pressure stages, as the large steps leading to a new pressure stage derive only from regulators. We see this separation in the example 4.2: no direct edge exists between an orange and a red node, and the same is true for all colour pairs. Thus, the pressure stages divide the gas network into different components.
In the following, when talking about a pressure stage, we mean the nodes belonging to this stage.

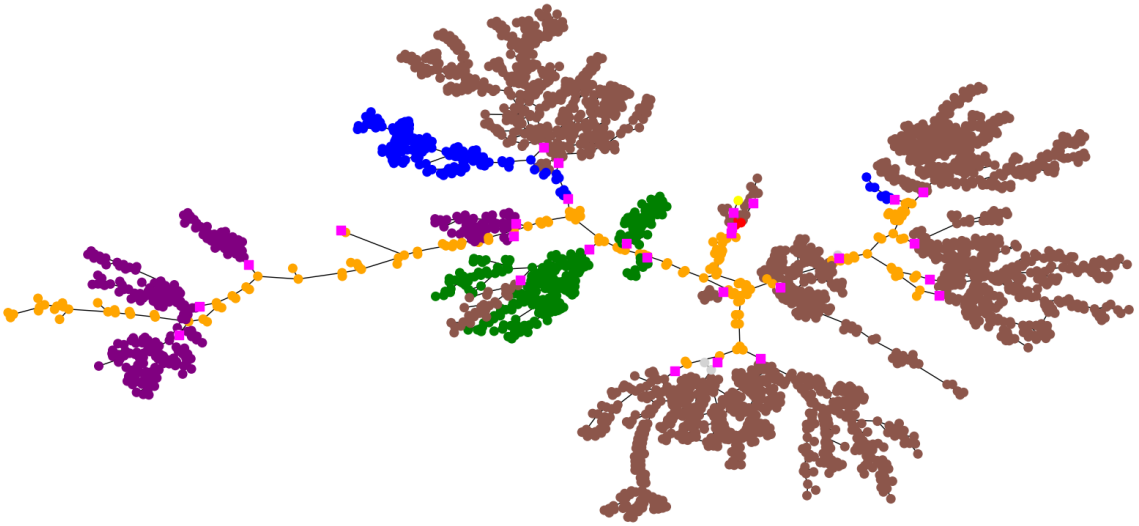


Figure 4.3.: Pressure stages and pressure areas. All brown nodes form one pressure stage, each connected component of brown nodes forms one pressure area.

When analysing the different pressure stages, we observe that not each pressure stage forms a connected graph. As an example, we look at another gas network shown in Figure 4.3. We observe that the orange pressure stage forms only one connected component, while the brown pressure stage consists of multiple components that are not directly connected among each other. For each pressure stage, these components are called the *pressure areas* of this stage (i.e., each component is a single pressure area). In the context of gas networks, this can mean, for example, that the orange pressure stage is a main supply line with higher pressure. From this main line, multiple branches branch off and lead (directly or with other stages between) to areas of the brown stage with lower pressure.

4.3.2. Parameters based on Pressure Stages

Our first parameter is the average pressure $Pr_{avr}(G)$ in the network G. To determine this, we do not use the pressure $p(v)$ of the nodes $v \in V$, but the pressure $p(e)$ of the edges $e \in E_p$. This allows us to weight the pressure values with the length of the pipelines to get a much more meaningful average pressure value. Thus, we define this parameter as follows:

$$\mathrm{Pr_{avr}(G)} = \frac{\sum_{e \in E_p} l(e) \cdot p(e)}{\sum_{e \in E_p} l(e)}. \tag{4.10}$$

In addition to that, we also compute the minimum and maximum pressure values, denoted as $Pr_{min}(G)$ and $Pr_{max}(G)$.
The next two parameters are the number of different pressure stages $NumStages(G)$ and the number of pressure areas $NumAreas(G)$.

4.3.3. Parameters based on Pressure Areas

We now take a look at the connections between the pressure stages and areas. For this we modify our graph in the following way. First, the graph is transformed to a multigraph. While this does not change the graph immediately, it allows multiple edges in the next step. In this next step, each pressure area is contracted to one single node. Each edge between those nodes stands for one connection between their corresponding pressure areas. If two areas are connected by multiple connections, this is represented by multiple edges between their contracted nodes. We refer to this modified multigraph as the *area connection graph*

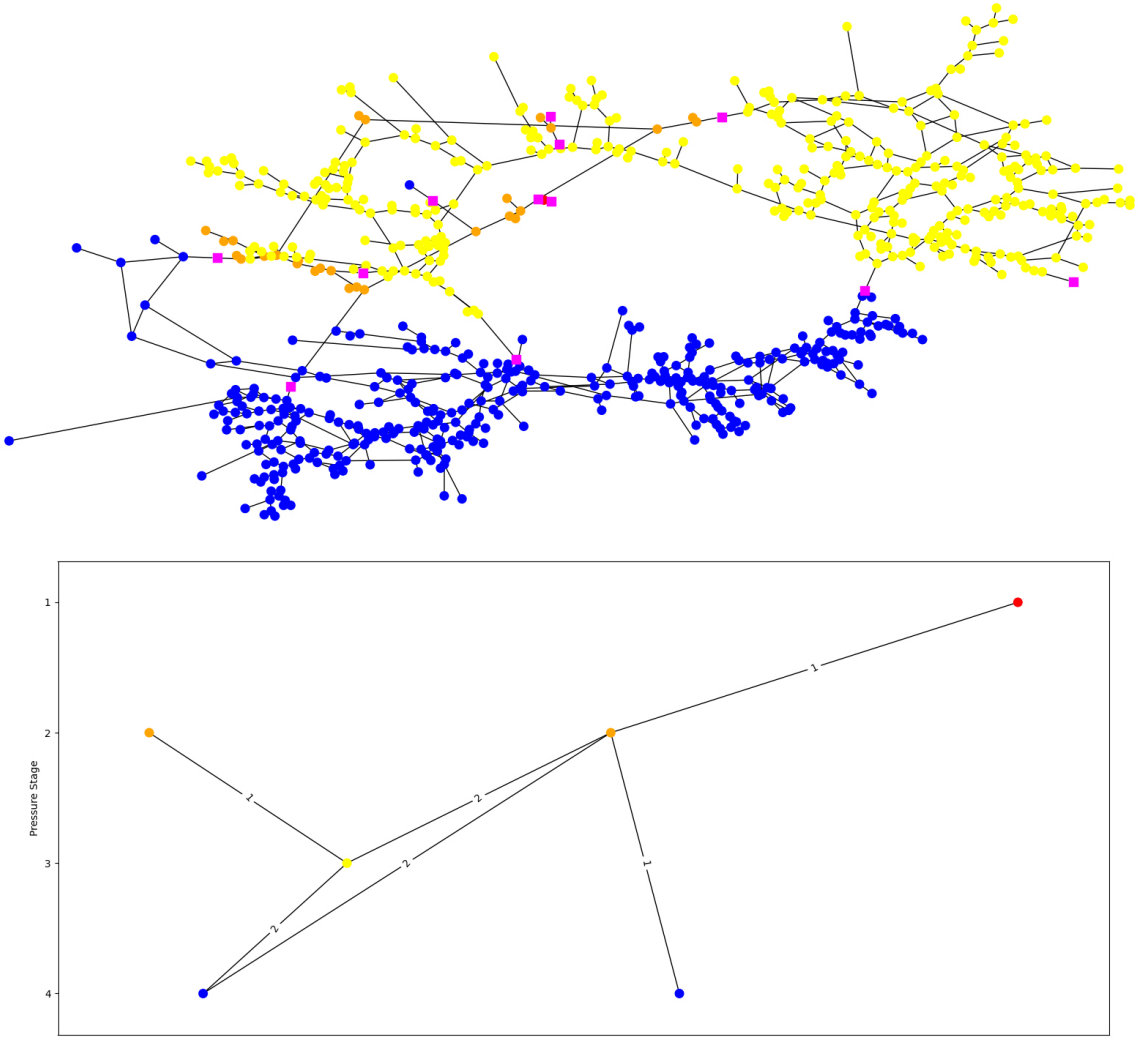


Figure 4.4.: Contraction of each pressure area to a single node. (a) Shows the original gas network, (b) shows the contracted area connection graph.

G_{ACG}. An example contraction is shown in Figure 4.4: (a) shows the gas network with its pressure stages and areas, (b) shows the contracted multigraph. Multiple edges are visualised as a single edge with the actual number of edges, which we call the *cardinality* of an edge, shown as the edge label. Take for example the large blue area. It is connected to the larger orange area by two regulators, therefore the number of edges (i.e., the edge label) between their contracted nodes is 2. The same blue area is also connected to the yellow area, again by two regulators. In the visualisation, the y-coordinate of a contracted node denotes the pressure stage the node belongs to. As before, the first pressure stage (shown at the top) is the one with the highest pressure value. When looking at a specific node, we refer to edges that come from a higher pressure stage node as incoming edges and to edges that go to lower pressure nodes as outgoing edges.
Note that from the visualisation of the area connection graph, one can also easily read the parameters *NumStages* (the height of the graph) and *NumAreas* (the number of nodes in G_{ACG}).

With the area connection graph, we examine several parameters that describe the connections between the pressure areas. The most straightforward parameter is the number of regulators in G. That is the number of connections between pressure areas and therefore the sum of all edge cardinalities in G_{ACG}. We denote this parameter as $NumRegulators(G)$.

The next two parameters are the number of source areas $NumSources(G)$ and the number of sink areas $NumSinks(G)$. A source area is a pressure area that is not connected to any area of a higher pressure stage, and a sink area is an area that is not connected to any area of a lower pressure stage. In G_{ACG}, these are the nodes with no incoming or no outgoing edge, respectively.

Next, we examine the ingoing edges of the areas that are not source areas. Each ingoing edge can be seen as a regulator that supplies such an area, coming from an area with higher pressure. We now examine by how many regulators an area is supplied and from how many different pressure areas and pressure stages these supplies come from. These examinations are performed on the area connection graph G_{ACG}. Recap that in this graph, each node represents one pressure area and each edge is associated with a cardinality.
First, we have the parameter $numSuppliesTotal(v)$ that denotes the number of regulators of higher pressure a given node v is supplied by. This corresponds to the sum of cardinalities of all ingoing edges of v. Next, we determine from how many different pressure areas these regulators come. This corresponds to the number of ingoing edges of v, and we refer to this parameter as $numSuppliesAreas(v)$. Lastly, we examine the number of different pressure stages the supplying pressure areas belong to. This parameter is called $numSuppliesStages(v)$. As an example, we again take a look at the left blue node in Figure 4.4 (b). That node is supplied by $2 + 2 = 4$ regulators, coming from two pressure areas that also come from two different stages. Of course, the two latter parameters are not always the same: the yellow node for example is supplied by two areas that have the same pressure stage. However, the inequality $numSuppliesTotal \geq numSuppliesAreas \geq numSuppliesStages$ always holds.
We compute these parameters for all nodes in G_{ACG} that do not represent a source area, and then compute for each parameter the maximum and the arithmetic mean among all these nodes.

Another aspect of the connections between pressure stages is that stages can be "skipped" by a regulator, i.e., the pressure is not regulated to the next pressure stage, but to an even lower stage. In the network shown in Figure 4.4 (b), this is the case for the right orange node. From this node, an edge goes to the blue pressure stage, skipping the yellow stage. For each edge in G_{ACG}, we compute how many stages are skipped. We then calculate the maximum and the arithmetic mean among all edges and refer to these parameters as $SkippedStages(G)$.

4.3.4. Expectations

Similar to the inner diameter of pipelines, the pressure is only an upper bound for the amount of gas that can be transferred through a network in a fixed amount of time. Thus, it does not provide much information about the actual amount of transferred gas, making it hard to predict parameters like the average pressure value upfront.
Still, we have some assumptions concerning the number of pressure stages and areas. Due to the structure of rural areas (that is, a long main supply line with several branches), we expect to have a higher number of pressure areas belonging to the same pressure stage compared to inner cities. In the latter, we expect to see less different pressure stages, but also more areas supplied by multiple inputs. Because of our expectation to observe more pressure areas and more pressure stages, we also presume the number of regulators to be higher in rural areas.

4.3.5. Examining previous Parameters on Lower Pressure Stages

When comparing the networks with the parameters from this chapter, this is always done on the complete graph. In addition, it can be of interest to compare certain parameters

only on the lower pressure stages that represent the gas network level of consumers. In this subsection, we first explain how we compute this so-called *low pressure graph* G_{low} and then describe which parameters we analyse on it.
To compute G_{low}, we have to decide which pressure stages it should consist of. Since some networks contain very few nodes on the lowest pressure stage, we do not find it sufficient to use only the nodes of the lowest stage. Instead, we want G_{low} to contain roughly 40 % of the nodes in G. To achieve this, we iterate through the pressure stages from lowest to highest and add all nodes of the current stage to G_{low} until it contains at least 20 % of all nodes in G. These 20 % are the minimum number of nodes we want to be in G_{low}. Next, we add more pressure stages to G_{low} until the proportion of nodes in G_{low} is as close to the wanted 40 % as possible. Note that we always add *all* nodes of the current pressure stage, which can result in G_{low} consisting of more than 40 % of the nodes in G. The reason we do not add pressure stages until at least 40 % of nodes are in G_{low} is that sometimes few pressure stages make of a large portion of the network. In some of these cases, adding the next stage to cross the 40 % threshold would result in more than 80 % of all nodes to belong to G_{low}, which is undesirable for our purposes.
Also note that G_{low} does not have to be a connected graph.

On this low pressure graph G_{low}, we examine three of the parameters described above, namely the pipeline lengths, the inner diameter of pipelines and the node degree. For each of these we examine the maximum and the arithmetic mean. We denote these parameters on G_{low} as Len_{low}, ID_{low} and Deg_{low}.
Our expectations regarding these parameters are quite similar to their counterparts on the whole network. However, we think that these expected characteristics are even more pronounced in the lower pressure stages.

4.4. 2-Core and Attached Trees

In this section, we explain several parameters that describe the 2-core G_2 and its complement $\overline{G_2}$ of a graph G. With these parameters, we try to describe the structure of G by splitting it into two major components. With the parameters examined on these components, a large part of the structure of the graph is captured.
In the following, we first explain what the 2-core and its complement are. Then, we go through the parameters we selected to describe both of them.

4.4.1. Explaining the 2-Core

The 2-core of a graph G is computed by iteratively removing nodes with degree less than 2. Since all of our graphs are connected, this simplifies to removing leaves of the graph, i.e., nodes with degree 1. Removing a node v means to remove it from V as well as removing all edges from E that are incident to v. At the end of this procedure, all nodes in the remaining graph G_2 have degree 2 or more. Note that for most graphs multiple iterations are required, as removing all leaves from a graph can create new leaves to be removed in the following iteration, and so on. The remaining nodes, i.e., the 2-core, are exactly the nodes that are either part of a cycle in G or are part of a path that connects two cycles. We refer to the set of nodes and the number of nodes in G_2 as V_2 and n_2.
As the complement graph $\overline{G_2}$ of the 2-core we denote the subgraph induced by all nodes that have been removed while computing the 2-core. These nodes have the characteristic that they form a so-called forest: a set of unconnected components where each component is a tree. This property follows directly from the observation that no cycles can exist in $\overline{G_2}$ as these would belong to G_2.
With these definitions, we can now describe G as the combination of two components: the 2-core and its complement that we will often refer to as the attached trees. In Figure 4.5

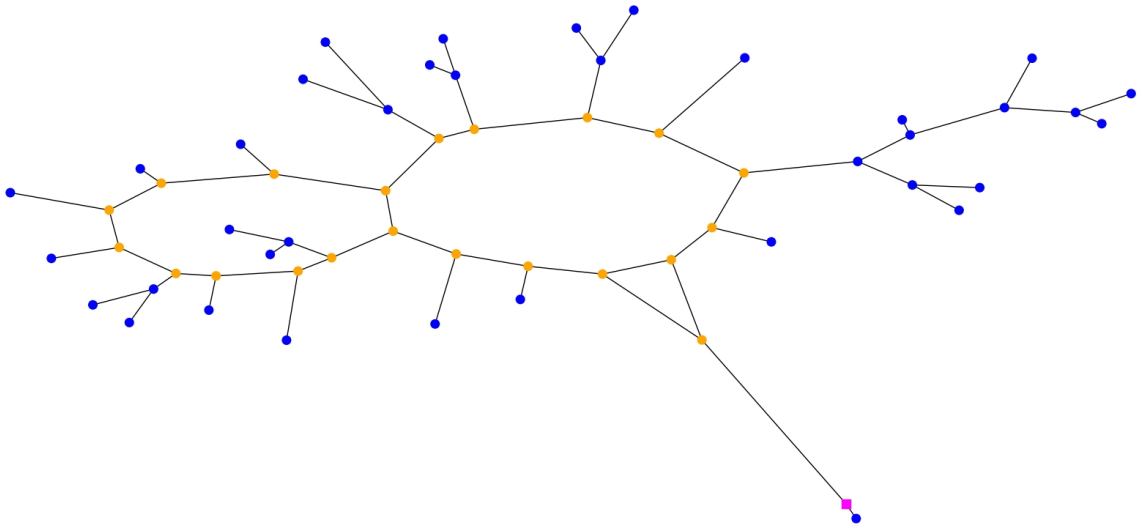


Figure 4.5.: Decomposition of a gas network into the 2-core and its complement. Orange nodes belong to the 2-core, the blue nodes are the attached trees in the complement. The pink square node represents a source of the network.

and Figure 4.6 the decompositions of two example gas networks into their 2-cores (orange nodes) and their complements (blue nodes) are shown. Again, sources of the network are highlighted as pink squares.
In the context of gas networks, we assume that the 2-core can be seen as the main supply infrastructure, while the attached trees are the supply lines to the consumers.

4.4.2. Structure of the 2-Core

In this subsection, we present parameters that describe the structure of the 2-core. This also includes information on how the trees are attached to the 2-core. The first parameter is the number of nodes that belong to G_2 in relation to the total number of nodes in G:

$$\text{SizeTwocore(G)} = \frac{n_2}{n}. \tag{4.11}$$

For the next parameters, we further decompose the 2-core. In the following, when we talk about the node degree, we refer to the degree in G_2 if not stated otherwise.
The 2-core is composed of nodes with degree ≥ 3 and nodes with degree 2. Consider two nodes u, v with $d(u) \geq 3, d(v) \geq 3$ that are connected by a path on which there are no other nodes with degree ≥ 3. We denote such a path as a *core path*. The number of nodes with degree 2 on a core path p is either 0 if p is only a single edge $\{u, v\}$, or greater zero if other nodes are on the path. The latter case is often seen in Figure 4.5, single edges are often seen in Figure 4.6. Each node on such a path is the "connection" point in G_2 to the root of at least one attached tree in $\overline{G_2}$. This follows from the property that in G no nodes with degree 2 exist, and therefore a node with degree 2 in G_2 has to have at least one edge incident to a node in $\overline{G_2}$. For each attached tree, we call this connection point the *origin* of the tree. A node in G_2 is an *origin node* if it is the origin of at least one tree. Therefore, the number of nodes on a core path p is a lower bound of the number of trees that are attached on p. As an example, we look at the leftmost core path of the graph in Figure 4.5: there are 8 nodes on the path, each of which is the origin of exactly one tree.
Given this, we define the next parameter $numOrigins(p)$ as the number of nodes with degree 2 on a core path p between two nodes with degree ≥ 3 in G_2. To evaluate this parameter on the whole network G, we compute $numOrigins(p)$ for each core path p

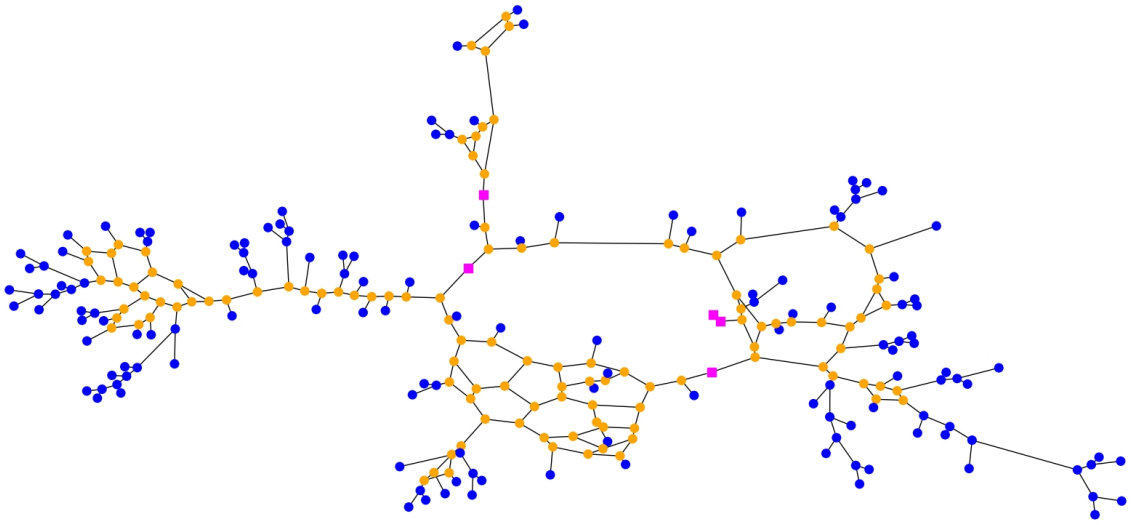


Figure 4.6.: Decomposition of a gas network into the 2-core and its complement. Orange nodes belong to the 2-core, the blue nodes are the attached trees in the complement. The pink square nodes represent either regulators or sources of the network.

and compute four statistical numbers: the minimum $NumOrigins_{min}(G)$ and maximum $NumOrigins_{max}(G)$, the arithmetic mean $NumOrigins_{avr}(G)$ and the standard deviation $NumOrigins_{stdev}(G)$.

When analysing these core paths between two nodes u and v, we are not only interested in the number of nodes on that path, but also in the distances between them. This distance is exactly the length of the edge between two nodes. For the first and the last node of the path we use the distance to u or v, respectively. As before, we analyse all these distances in G_2 and compute the minimum $DistancesOrigins_{min}(G)$ and maximum $DistancesOrigins_{max}(G)$, the arithmetic mean $DistancesOrigins_{avr}(G)$ and the standard deviation $DistancesOrigins_{stdev}(G)$.

When introducing the concept of core paths, we distinguished between nodes with degree ≥ 3 and nodes with degree 2 in G_2. The ratio between the number of both node types can be of interest, as we already see significant differences in the two examples mentioned earlier. Thus, we introduce the parameter $RatioHigherDegree(G)$ that is computed as the proportion of nodes in G_2 that have degree ≥ 3:

$$\text{RatioHigherDegree(G)} = \frac{|\{v \in V_2 \mid d(v) \geq 3\}|}{n_2}. \tag{4.12}$$

The last two parameters of this subsection cover further aspects of the origin nodes. Our assumption is that most time, origin nodes have degree 2 in G_2. Furthermore, it seems to be rare that one node is the origin for more than one attached tree. For both examples shown previously, one can quickly confirm that both assumptions are true. But, as we see in cut-outs of another gas network in Figure 4.7 and in Figure 4.8, both of these events can indeed occur. The number of origin nodes with degree ≥ 3 in G_2 and the number of origin nodes that are origin of multiple trees are called $OriginHigherDegree(G)$ and $OriginsMultiple(G)$, respectively.

4.4.3. Structure of the Complement Graph

In the last subsection, we found parameters that describe patterns of *where* the attached trees of $\overline{G_2}$ are placed. In this subsection, we introduce parameters that characterise the

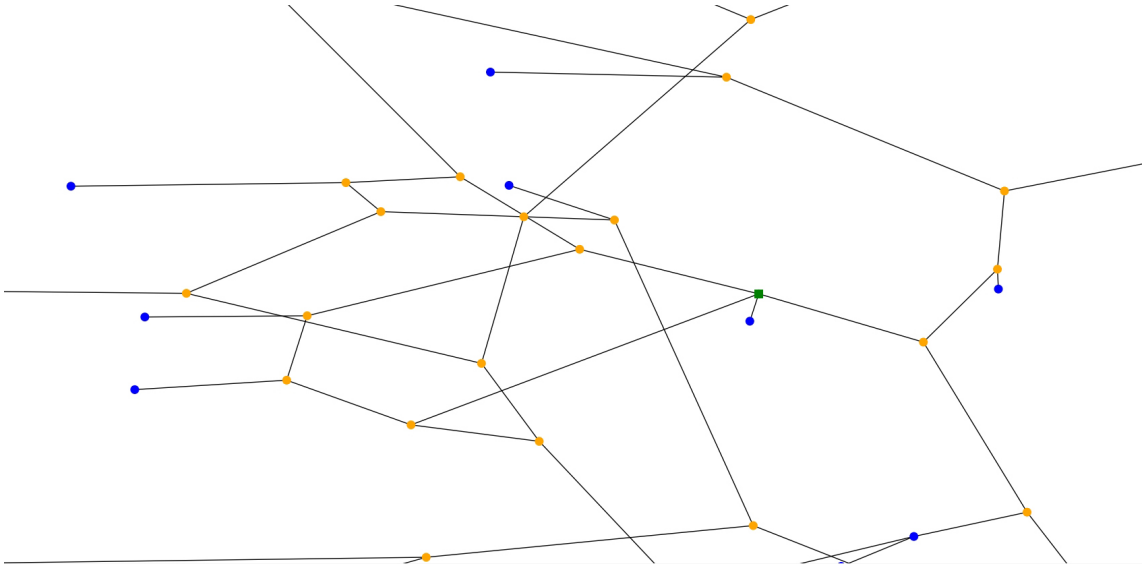


Figure 4.7.: Example of an origin node, namely the green square one, that has degree 3 in G_2.

structure of the trees themselves. We analyse three characteristics of each tree t: the *size*, the *depth* and the *number of leaves*. In Figure 4.9 an example of an attached tree is shown, including its origin node in the 2-core. The edges are labelled with their length. For the following parameters, we will refer to this example tree.

The size $size(t)$ of a tree is simply the number of nodes of t. In the example tree, $size(t) = 13$.
The depth $depth(t)$ of a tree is the maximum length of all shortest paths from the root r to each leaf, i.e., it is the distance from r to the furthest leaf. Thus, it is a way of measuring how large t is in terms of real-world distance. In our example, the leaf with the longest shortest path to the root is v, and this results in $depth(t) = 215.0$.
The last parameter is the number of leaves of t, denoted with $numLeaves(t)$. In the example tree, $numLeaves(t) = 6$.
Note that a tree t may only consist of a single node. In that case, we define $size(t) = 1$, $depth(t) = 0$ and $numLeaves(t) = 1$.

These three parameters are computed for each tree in $\overline{G_2}$. The corresponding global parameters that are computed over all trees are denoted as $SizeTrees(G)$, $DepthTrees(G)$ and $LeavesTrees(G)$. For each of these four parameters we analyse the minimum, the maximum, the arithmetic mean and the standard deviation.

4.4.4. Expectations

For most of the parameters based on G_2 and $\overline{G_2}$, we find it hard to predict specific patterns. In general, our hope is that the description of the structure of both components shows common features and differences between instances that other parameters do not cover.
However, for some parameters we have some expectations. We assume the distances between attached trees to be higher in rural areas and industrial areas than in residential areas and inner cities since the density of consumers is much higher in the latter. Because of the assumed larger meshedness in inner cities, we also assume the proportion of nodes in G_2 to be higher there compared to rural areas.

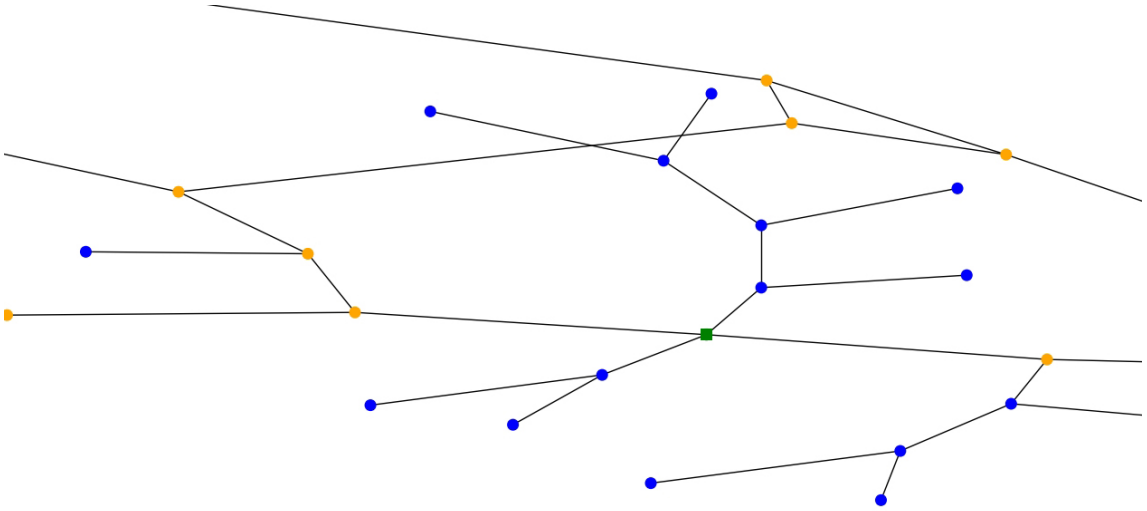


Figure 4.8.: Example of an origin node, namely the green square one, that is origin of two attached trees.

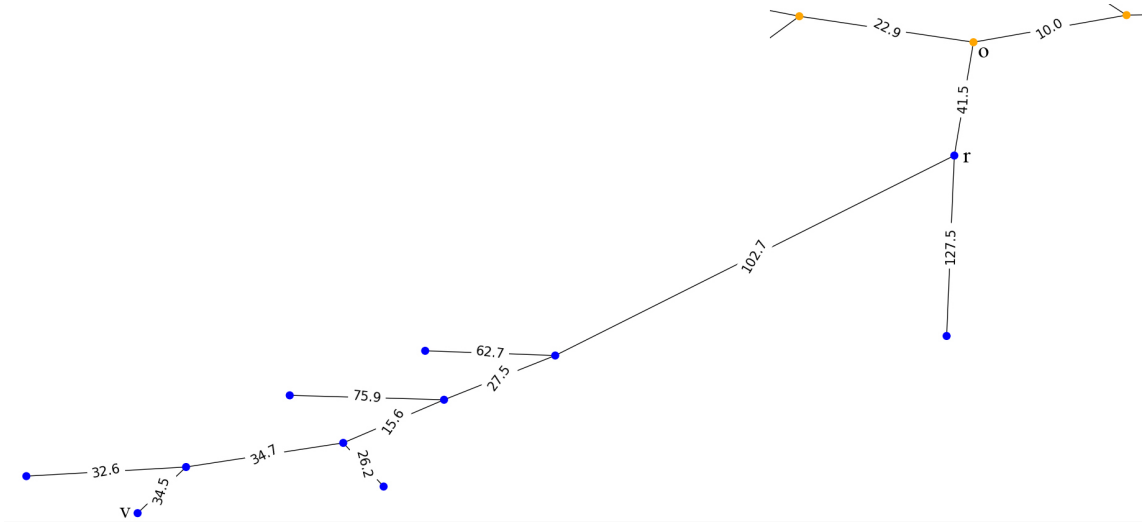


Figure 4.9.: Example of an attached tree. Three important nodes are the origin node o, the root r and the furthest node v.

4.5. Further Parameters

In this section, we discuss two more parameters that do not fit well in any of the previous sections.

The first of these is the treewidth of a graph G, denoted as $TW(G)$. This parameter describes how far G is from being a tree. The treewidth of trees is exactly 1, and the larger $TW(G)$ is, the further is G from being a tree. We will not discuss how the treewidth is being computed in this work, but it is worth mentioning that computing the treewidth is $\mathcal{NP}$-complete [ACP87]. Thus, we will use heuristics to approximate this parameter. Two well-known heuristics are the *Minimum Degree Heuristic* and the *Minimum Fill-in Heuristic*. As both of these overestimate the actual treewidth, we define $TW(G)$ as the minimum of the treewidths computed by the heuristics.
We expect the networks in more densely populated areas to be more meshed and thus less tree-like than in rural areas. Therefore, we expect larger treewidth values in these more populated networks.

The second parameter *IsPlanar(G)* is Boolean and states if a graph G is planar or not. A planar graph can be drawn in the plane in a way that no edges cross each other. In the context of gas networks, crossing edges mean that two pipelines at different depths pass through the same point in the plane. In planar networks, this does never happen. Recall that our given network instances have no coordinates; thus, if the graph is shown to be planar, it does not necessarily mean that no pipelines cross in the real network. It just means that the pipelines *can* be laid in a way that they do not cross each other. On the other hand, if the graph is shown to be not planar, this means that also in the real network at least two pipelines cross.
Due to the fact that rural areas provide more space for the pipelines to be laid and the connected points are less densely distributed than in inner cities, our expectation is that rural areas tend to be planar more often than inner cities.

5. Parameter Analysis

In this chapter, we analyse the parameters described in the previous chapter. For each category of parameters (see the overview in Table 4.1), we first present a tabular overview of the computed values in all gas network instances. We then analyse and evaluate these values, i.e., point out differences and common features between the instances. As stated before, we both compare the instances within a class and between the classes to find out whether the class label set seems to be meaningful or not. Thus, for each parameter there are several possible outcomes. If it has similar values for all networks of each class, but dissimilar values across the classes, this is an indicator that the assumed classes are correct. On the other hand, if the parameter has similar values for all networks of two classes, it is an indicator that these classes may not be distinguishable and should be one class. And, of course, a parameter can have no meaning at all, for example if the values differ on each instance significantly, no matter of which class the instances are.
We do not expect all parameters to have the same outcome. On the contrary, we expect all of the above outcomes to occur. Therefore, in Section 5.6, we bring together our findings and do an overall evaluation, where we come to a conclusion regarding what our class set may look like.

Our data set consists of 11 gas network instances. In Table 5.1, these instances are shown together with their number of nodes and edges and their label. Remember that these labels are not fixed but denote which type of area the instances originate from. The instances are named with the first three letters of their label, followed by a numeric identifier. The visualisations of all instances are shown in Appendix A.1 as well as the visualisations of their pressure stages in Appendix A.2.

5.1. Length, Geographic Extent and Inner Diameter

5.1.1. Analysis

We begin our analysis with the parameters length and inner diameter, shown in Table 5.2. The maximum length shows quite similar values between the old town and the residential areas, although Res3 is higher than the others. In contrast, all rural areas and inner cities have significantly higher values. Among these networks, the values are widely spread within similar intervals. The industrial area networks both have quite similar values and differ from the other networks, ranking between the before-mentioned ranges. The average and median length show that the old town network and the residential areas again have

Table 5.1.: Overview of the gas network instances.

Gas Network Instance	Attribute		
	Label	Nodes	Edges
Inn1	Inner City	741	883
Inn2	Inner City	2174	2666
Old1	Old Town	166	201
Res1	Residential Area	258	283
Res2	Residential Area	438	495
Res3	Residential Area	391	422
Rur1	Rural Area	219	234
Rur2	Rural Area	2782	2967
Rur3	Rural Area	904	988
Ind1	Industrial Area	59	61
Ind2	Industrial Area	52	55

Table 5.2.: Length and inner diameter.

Gas Network Instance	Parameter							
	Len_{min}	Len_{max}	Len_{avr}	Len_{mdn}	ID_{min}	ID_{max}	ID_{avr}	ID_{mdn}
Inn1	0.3	5499	131.5	86.7	51.4	900	141.6	107.1
Inn2	0.1	2272	99.1	67.3	8.0	546	177.9	160.3
Old1	0.3	528	69.7	49.2	32.6	250	119.9	107.9
Res1	0.5	520	77.2	55.2	37.2	316	111.4	102.2
Res2	0.1	730	85.7	52.6	26.2	210	118.4	107.9
Res3	0.4	1221	75.7	49.0	26.2	900	107.8	53.1
Rur1	0.2	3046	105.0	47.1	26.2	159	88.7	90.0
Rur2	0.4	5678	96.8	49.7	20.4	900	97.8	51.4
Rur3	0.3	2418	114.6	70.7	26.2	900	86.2	51.4
Ind1	0.9	1884	167.4	129.9	40.8	256	133.7	130.8
Ind2	0.9	1628	275.9	186.3	51.4	900	185.1	121.9

similar values that are lower than the ones of the rural areas and the inner cities. The most noticeable observation here is that the values of the industrial areas are much larger, especially for the median. The minimum length does not seem to provide much information, only the industrial areas have mentionable higher values than the rest.

For the inner diameter the minimum and maximum do not yield useful information. The arithmetic mean in contrast is quite interesting, as it shows non-overlapping intervals between the inner cities, the residential areas plus the old town, and the rural areas. However, the gap between Rur2 and Res3 is not that large. Both industrial networks have values quite similar to both inner cities. Meanwhile, the median of the inner diameter does not provide such useful information.

Next, we analyse the parameters that describe the geographic extent and centrality of the networks. These are shown in Table 5.3. The absolute values ($PL_{avr}, Diam, Rad, Cent$) all seem to strongly correlate with the number of nodes of the network. Thus, we observe differences between the networks, but they do not contain much new information. We therefore focus on the parameters that we scaled by the number of nodes. For these four

Table 5.3.: Average path length, diameter, radius and centrality.

Gas Network Instance	Parameter							
	PL_{avr}	PL_{avr_n}	Diam	$Diam_n$	Rad	Rad_n	Cent	$Cent_n$
Inn1	2723	3.67	9473	12.78	5135	6.93	1946	2.63
Inn2	3965	1.82	14 818	6.82	7438	3.42	2744	1.26
Old1	508	3.06	1654	9.97	871	5.25	363	2.19
Res1	1181	4.58	3763	14.59	1886	7.31	819	3.17
Res2	1550	3.54	4785	10.92	2458	5.61	1013	2.31
Res3	1411	3.61	4402	11.26	2224	5.69	928	2.37
Rur1	3076	14.1	8852	40.42	4793	21.89	2191	10.00
Rur2	16 737	6.02	54 945	19.75	27 721	9.97	11 301	4.06
Rur3	7974	8.82	17 771	19.66	10 433	11.54	6000	6.64
Ind1	1070	18.14	3611	61.20	1885	31.94	767	13.00
Ind2	1679	32.29	5570	107.11	2955	56.83	1175	22.60

parameters we observe some interesting common features and differences between the instances. The inner city, the old town and the residential area networks have quite similar values, though the residential area ones tend to be a bit higher. In contrast, the values of the rural areas and even more those of the industrial areas stand out clearly from the other networks as they are considerably greater. This seems to confirm our assumption that the node density is considerably smaller in rural areas and industrial areas. It also fits our previous observation that the average pipe lengths, and therefore the distances between two nodes, is greatest in industrial areas.

5.1.2. Evaluation

In general, the length and inner diameter indicate that the assumed class label set fits quite well, since many parameters have quite similar values within one class, but different intervals between the classes. An exception is the old town network that is very similar to the residential areas. With the maximum length, we can distinguish between the residential areas (plus the old town) and the other classes. The average and median length are useful to identify industrial areas, while the average diameter can be helpful to separate the inner cities from the rural areas.
Still, we also observe some large gaps between parameter values of the same network type, especially when looking at the two inner cities and the two industrial areas. Also, some intervals overlap, for example the average length of the inner cities and the rural areas.
The extent and centrality parameters show quite similar values in a comparatively small interval across the inner cities, residential areas and the old town. They are therefore not useful to distinguish between networks of these classes. However, they are in fact very helpful to separate the rural areas and industrial areas from the other types, and even to distinguish between these two classes.

5.2. Meshedness and Connectivity

5.2.1. Analysis

We now analyse the parameters that deal with the meshedness, these are shown in Table 5.4. The maximum node degree does not vary too much among all networks. Inn1 is the only network where $Deg_{max} = 5$, for all other classes with the exception of the industrial areas

Table 5.4.: Node degree and clustering coefficient.

Gas Network Instance	Parameter			
	Deg_{min}	Deg_{max}	Deg_{avr}	CC
Inn1	1	5	2.383	0.014
Inn2	1	4	2.453	0.017
Old1	1	3	2.422	0.000
Res1	1	3	2.194	0.012
Res2	1	4	2.260	0.008
Res3	1	4	2.159	0.005
Rur1	1	3	2.137	0.014
Rur2	1	4	2.133	0.005
Rur3	1	4	2.186	0.014
Ind1	1	3	2.068	0.017
Ind2	1	3	2.115	0.019

we observe instances with maximum degrees of both 3 and 4. In contrast, the average node degree offers some valuable information as the values are significantly greater in the inner cities and the old town in comparison to the other networks. Also, the values in these two classes are fairly close to each other. The residential areas and the rural areas have quite similar values, while those of the industrial areas are a bit smaller, especially in Ind1.
The clustering coefficient has the greatest values in the industrial area and the inner city networks. The values of the rural area and residential area networks are very similar to each other, while the old town has a value of zero.

5.2.2. Evaluation

The average node degree seems to be very useful to separate inner cities and old towns from the other classes. This matches our expectations that inner cities are more tightly meshed than the other networks. Another important observation is that the average node degree of the old town network is very similar to the inner cities. This is in contrast to the findings of the previous section, where the old town was very similar to the residential areas.
The average node degree does not differ too much among the residential areas and rural areas, even though it tends to be a bit smaller in the latter. Industrial areas seem to be the least meshed, though the average node degree interval is not that far away from the residential areas and rural areas. Thus, this parameter is an indicator, but not perfectly suited to differentiate between these three classes.
The clustering coefficient does not provide too much information. However, it is worth mentioning that it equals zero in the old town network. This is somewhat surprising, as the old town has a very similar average node degree as the inner cities, for which the clustering coefficient is significantly greater.

5.3. Pressure

5.3.1. Analysis

In this section, we analyse the parameters concerned with the pressure of the networks, starting with the parameters in Table 5.5. We observe that the minimum and maximum pressure do not provide much useful information as the values differ very much among

Table 5.5.: Minimum and maximum pressure, average pressure, number of pressure stages and number of pressure areas.

Gas Network Instance	Parameter				
	Pr_{min}	Pr_{max}	Pr_{avr}	NumStages	NumAreas
Inn1	0.0219	40.00	2.525	4	6
Inn2	0.0237	70.00	1.185	7	10
Old1	0.0236	0.54	0.127	2	3
Res1	0.0235	0.54	0.270	3	3
Res2	0.0228	40.00	0.299	8	10
Res3	0.2742	12.39	0.385	2	3
Rur1	0.0219	40.00	0.740	5	8
Rur2	0.1000	12.50	2.360	8	22
Rur3	0.4898	12.60	1.873	3	4
Ind1	0.5179	7.85	0.528	2	2
Ind2	3.4311	12.60	3.521	2	2

most networks. When it comes to the average pressure, we observe very homogenous, small values for the residential areas. The old town has an even smaller, but still quite similar value. For the rural areas and industrial areas, the values have a wide range and differ greatly from each other. Still, they all are greater than the residential area values. The average pressures of both inner cities are also significantly higher than the residential area ones, but also differ by factor 2 from each other.
Regarding the number of different pressure stages and pressure areas, we are interested in both the individual numbers and the difference between them. We observe that both industrial areas only have two stages and two areas. The classes inner city, residential area and rural area each have instances with both smaller and larger numbers. For the residential areas, Res2 stands out a lot from the other instances and looks more similar to the inner cities. It is also noticeable that the rural areas and the inner cities all have more areas than stages, especially Rur2 has an extremely high number of areas.

We observe similar results for the parameters shown in Table 5.6. The industrial areas are identical to each other, while the other classes have varying values. Again, Res2 and Rur2 are somewhat outstanding, both having a lot of regulators and skipped stages, quite similar as Inn2. The numbers of source and sink areas do not provide much information, although the number of sinks of Rur2 is extremely high.

We also analysed the connections between the pressure areas, the resulting parameter values are shown in Table 5.7. However, we have to state that these parameters do not provide new information; just as the previous parameters, the values vary quite a lot from each other, both within and between the classes. The only exceptions are again the two industrial areas. One parameter that is somewhat interesting is the average regulator supply, i.e., the number of regulators that supply an area. For this parameter, inner cities, residential areas and the old town tend to have higher numbers than the rural areas. Furthermore, Res2 is again more similar to the inner cities than to the other residential areas.

Lastly, we take a look at the low pressure graph parameters, shown in Table 5.8. We also compare them with the corresponding parameters of the full networks, shown in Table 5.2 (length and inner diameter) and Table 5.4 (node degree). For the maximum and the average pipeline length, we see that the inner city networks and two of the rural areas have significantly lower values than on the full graph, while the industrial areas did not

Table 5.6.: NumRegulators, NumSources, NumSinks, $\text{SkippedStages}_{\text{max}}$, $\text{SkippedStages}_{\text{avr}}$.

Gas Network Instance	**Parameter**				
	Regulators	Sources	Sinks	$\text{skipped}_{\text{max}}$	$\text{skipped}_{\text{avr}}$
Inn1	9	2	2	1	0.33
Inn2	25	2	5	2	1.16
Old1	4	2	1	0	0.00
Res1	3	1	2	1	0.33
Res2	20	3	3	3	1.25
Res3	2	2	1	0	0.00
Rur1	7	2	6	1	0.57
Rur2	27	1	17	5	2.78
Rur3	3	2	1	0	0.00
Ind1	1	1	1	0	0.00
Ind2	1	1	1	0	0.00

Table 5.7.: NumSuppliesTotal (ST), NumSuppliesAreas (SA), NumSuppliesStages (SS).

Gas Network Instance	**Parameter**					
	ST_{max}	ST_{avr}	SA_{max}	SA_{avr}	SS_{max}	SS_{avr}
Inn1	4	2.25	2	1.50	2	1.25
Inn2	13	3.13	2	1.13	1	1.00
Old1	4	4.00	2	2.00	1	1.00
Res1	2	1.50	1	1.00	1	1.00
Res2	10	2.86	4	1.57	3	1.43
Res3	2	2.00	2	2.00	1	1.00
Rur1	2	1.17	1	1.00	1	1.00
Rur2	3	1.29	2	1.05	2	1.05
Rur3	2	1.50	2	1.50	1	1.00
Ind1	1	1.00	1	1.00	1	1.00
Ind2	1	1.00	1	1.00	1	1.00

change at all. The old town and residential areas did slightly decrease, but not as much as the rural areas and inner cities. It stands out that the maximum length of Res3, which has also the greatest maximum length among the full residential area networks, did not change. Comparing the network types shows that the differences between inner cities and residential areas have decreased. Also, the rural areas and the residential areas are now closer to each other than on the full graph. The industrial areas still have much higher values.

As for the inner diameters, we observe that both the maximum and average are quite the same as on the full graph. The only exception is the average inner diameter of Inn1 that increased from 141.6 to 168.7 and is now very close to the value of Inn2. This also increased the gap between the inner cities and the residential areas.

The maximum node degree decreased from 5 to 4 in Inn1 while staying the same in all other networks. When comparing the average node degree on the low pressure graph with the full graph, we find both decreases and increases depending on the instance. However, these changes are mostly very small, and therefore, the analyses here are quite the same as on the full graph.

Table 5.8.: Low pressure graph: length, inner diameter, node degree.

Gas Network Instance	Parameter					
	Len_{max}	Len_{avr}	ID_{max}	ID_{avr}	Deg_{max}	Deg_{avr}
Inn1	822	102.0	900	168.7	4	2.438
Inn2	1365	88.3	546	170.3	4	2.478
Old1	355	61.8	250	119.3	3	2.458
Res1	520	68.1	184	107.1	3	2.217
Res2	544	78.6	210	120.9	4	2.248
Res3	1221	75.8	900	107.8	4	2.160
Rur1	385	71.3	131	91.1	3	2.055
Rur2	821	67.7	900	90.1	4	2.135
Rur3	496	89.7	900	85.9	4	2.255
Ind1	1884	168.4	256	133.7	3	2.069
Ind2	1628	279.4	900	184.2	3	2.120

5.3.2. Evaluation

We observed that the average pressure is indeed useful for classification. All residential areas as well as the old town have quite similar and significantly smaller average pressures than all other network classes. However, the other classes do not have very homogenous values for this parameter and also have similar intervals. Thus, the average pressure is well suited only to separate the residential areas and old towns from the other networks.
In contrast, the parameters concerning the different pressure stages and pressure areas do not show characteristics that are clear enough to use them for classification. Most parameters vary a lot among the instances of the classes, and also among the different classes there are a lot of overlapping intervals. One interesting observation is that the instance Res2 tends to be more similar to the inner city instances than to the other residential area instances. Although this is only true for a subset of the parameters discussed here, it suggests that the borders between inner cities and residential areas may be blurred with respect to these parameters.
The low pressure graph does not provide too much additional information. The average inner diameter values are slightly clearer here since the inner cities are closer together and the difference to the residential areas does increase. Therefore, this parameter can be used in addition to its full graph counterpart. Furthermore, the intervals for the maximum and average length are closer together on the low pressure graph. This may indicate that the differences in length between inner cities and rural areas on the one side and the residential areas and old town on the other side are mainly located on the higher pressure stages.

5.4. 2-Core and Attached Trees

5.4.1. Analysis

In this subsection, we analyse the 2-core G_2 and its complement $\overline{G_2}$, starting with the parameters presented in Table 5.9. For the proportion of nodes in G_2, we observe the highest values in the old town and the inner city networks. These have significantly higher values than all other networks, with the exception of Res2 that is ranked between the former and both other residential areas. The other networks do not differ much from each other.
We observe very similar results for the ratio of nodes in G_2 that have degree ≥ 3, but

Table 5.9.: SizeTwocore (STC), RatioHigherDegree (RHD), OriginsHigherDegree (OHD), OriginsMultiple (OM).

Gas Network Instance	**Parameter**			
	STC	RHD	OHD	OM
Inn1	0.601	0.616	9	8
Inn2	0.653	0.654	13	6
Old1	0.663	0.636	0	0
Res1	0.442	0.439	0	0
Res2	0.567	0.460	1	0
Res3	0.419	0.372	0	0
Rur1	0.411	0.333	0	0
Rur2	0.373	0.348	4	6
Rur3	0.433	0.427	1	8
Ind1	0.356	0.191	0	0
Ind2	0.385	0.300	0	0

here Res2 is fairly close to the other residential areas, making the difference between those and the inner cities and the old town noticeably larger. The two industrial areas have the lowest values here, though the larger one is not much smaller than some rural areas and residential areas.
The next parameters deal with the origin nodes in G_2. Both the number of origin nodes with degree ≥ 3 and the number of origin nodes that are the origin of more than one tree show pretty interesting results. We observe high numbers for both inner cities, while both numbers are zero for the old town, both industrial areas and two of the residential areas. Also, Res2 has zero multiple origin nodes and only one with degree ≥ 3. For the rural areas we see varying values, ranging from zero to values similar to the inner cities.

We proceed with the statistical parameters concerned with the number of origin nodes on core paths and the distances between them (Table 5.10). For the maximum number we observe very heterogenous values, both within and between the classes. The average though shows noticeably smaller values for the inner cities and the old town and larger values for the industrial areas. The rural areas and residential areas have quite similar values.
Looking at the distances between origin nodes, we see that the inner cities and the rural areas have similar maximum values, which are significantly higher than the ones of the old town, the residential areas and the industrial areas. The latter three also have quite similar values, though Res3 has a noticeably higher maximum. The average distance is significantly higher in rural areas and industrial areas compared to residential areas and the old town. The two inner city networks differ strongly from each other: Inn2 has values similar to the residential areas, Inn1 looks similar to the rural areas.
For both parameters the standard deviation mostly correlates with the average values and therefore does not offer much additional information. An exception to this is DO_{stdev} in the industrial area networks as it is much smaller than for the rural areas, despite having larger average distances. It also stands out that the maximum distances in the industrial areas are very small compared to their average.

At last, we analyse $\overline{G_2}$, i.e., the attached trees. The statistics about the size and the depth of the trees are presented in Table 5.11, the statistics about the number of leaves are shown in Table 5.12.
The maximum sizes are very inconsistent and therefore do not offer much valuable infor-

Table 5.10.: NumOrigins (NO), DistancesOrigins (DO).

Gas Network Instance	Parameter							
	NO_{min}	NO_{max}	NO_{avr}	NO_{stdev}	DO_{min}	DO_{max}	DO_{avr}	DO_{stdev}
Inn1	0	7	0.397	0.882	0.4	5499	184.0	501.0
Inn2	0	19	0.335	1.073	0.9	2272	112.8	162.5
Old1	0	5	0.352	0.717	0.8	528	84.9	96.0
Res1	0	10	0.813	1.581	0.5	520	82.8	98.4
Res2	0	6	0.690	1.125	0.7	730	108.5	124.9
Res3	0	7	1.120	1.601	0.5	1221	96.5	145.0
Rur1	0	5	1.222	1.489	0.9	3046	186.4	398.9
Rur2	0	16	1.205	2.032	0.4	3435	140.6	287.5
Rur3	0	17	0.889	1.797	0.3	2398	142.8	250.5
Ind1	0	8	2.833	3.078	1.1	526	179.7	136.6
Ind2	0	4	1.556	1.423	15.1	517	239.1	148.8

mation. The average size separates the instances quite clearly into two groups: both inner cities, the old town and Res2 on the one side, the other networks on the other side. The standard deviation is again mostly correlated with the average values, but interestingly this is not true for Old1 and Res2. On these instances, that parameter is much smaller than on the two inner cities that have similar average values.
The depth parameters do not provide any information at all, as both the maximum and the average values are widely spread among all instances and the intervals of the distinct classes overlap heavily.
For the maximum number of leaves, we also observe widely spread values, though they tend to be higher in rural areas and smaller in industrial areas. The average number of leaves shows results that are quite similar to the average size of the trees: we observe similar values among the inner cities, the old town and Res2 on the one hand, and similar values among the remaining instances on the other hand. The gap between these intervals is not too large, but still clearly existing. Additionally, the standard deviation is again noticeably smaller for Res2 and Old1 than for the two inner cities. Furthermore, the standard deviation is significantly higher for all rural areas compared to the other networks.

5.4.2. Evaluation

The proportion of nodes that belong to the 2-core seems to be a quite useful parameter, as it separates the inner cities plus the old town from the other classes. Like for the meshedness parameters, the old town network is again very similar to the inner cities. Interestingly, Res2 again stands out from the other two residential areas and seems to be located between these and the inner cities. Looking back at the average node degree, Res2 was more similar to the residential areas, but among these instances it featured the smallest difference to the inner cities. This may indicate a correlation between the meshedness and the size of the 2-core. This correlation seems logical, since a higher meshedness tends to lead to more nodes being part of cycles and therefore belonging to G_2.
The ratio of nodes with higher degree also seems useful, again for identifying both inner cities and old towns. Compared to the 2-core size it has the advantage that Res2 is clearly similar to the other networks instead of being located between the inner cities and the residential areas.
Also quite interesting are the origin nodes parameters because the old town network is

Table 5.11.: Attached trees: SizeTrees (S) and DepthTrees (D).

Gas Network Instance	Parameter							
	S_{min}	S_{max}	S_{avr}	S_{stdev}	D_{min}	D_{max}	D_{avr}	D_{stdev}
Inn1	1	17	1.672	2.082	0	822	38.2	120.5
Inn2	1	31	1.535	2.207	0	4681	31.4	234.7
Old1	1	5	1.514	1.130	0	273	19.4	53.9
Res1	1	13	2.361	2.698	0	1043	61.0	156.7
Res2	1	8	1.597	1.410	0	359	22.7	62.4
Res3	1	27	2.204	3.278	0	615	31.4	85.5
Rur1	1	27	2.346	4.510	0	903	58.5	181.7
Rur2	1	70	2.617	5.798	0	9891	77.5	528.5
Rur3	1	54	2.250	4.492	0	2521	66.0	258.8
Ind1	1	11	2.235	2.365	0	358	66.5	119.6
Ind2	1	9	2.286	2.575	0	1721	304.1	592.2

Table 5.12.: Attached trees: LeavesTrees (Leaves).

Gas Network Instance	Parameter			
	$Leaves_{min}$	$Leaves_{max}$	$Leaves_{avr}$	$Leaves_{stdev}$
Inn1	1	9	1.328	1.039
Inn2	1	16	1.260	1.088
Old1	1	3	1.243	0.541
Res1	1	7	1.672	1.352
Res2	1	4	1.277	0.685
Res3	1	14	1.583	1.634
Rur1	1	14	1.636	2.219
Rur2	1	35	1.803	2.889
Rur3	1	30	1.640	2.385
Ind1	1	6	1.588	1.191
Ind2	1	5	1.571	1.237

clearly different to both inner city networks. Thus, they can be used to distinguish between these classes that are very similar with respect to the previous two parameters.

Even though most of the parameters dealing with the origin nodes do not provide too much additional information, some of them can still be used for classification. The average number of origin nodes on core paths seems helpful to separate the inner cities and the old town from the other networks. The average distances between origin nodes can be used to distinguish between the residential areas plus the old town and the rural areas plus the industrial areas. An interesting observation concerning the distances is that the industrial areas exhibit the largest average distances, but very small maximum values and standard deviations. Thus, these distances may be helpful to identify industrial areas.

The attached trees provide some useful information in terms of their average size and number of leaves, as these values separate the networks clearly into two sides. As has sometimes been the case before, Old1 and also Res2 are similar to both inner cities with regard to these averages. Interestingly, they can be distinguished from the inner cities with the standard deviations. The standard deviation of the number of leaves can also be used

Table 5.13.: Treewidth and planarity.

Gas Network Instance	Parameter	
	Treewidth	IsPlanar
Inn1	8	No
Inn2	16	No
Old1	7	No
Res1	4	Yes
Res2	6	No
Res3	5	Yes
Rur1	3	Yes
Rur2	7	No
Rur3	4	Yes
Ind1	2	Yes
Ind2	3	Yes

to identify rural areas, as these have significantly higher values than the other instances. In contrast, the depth does not provide any information that can be used for classification.

5.5. Further Parameters

5.5.1. Analysis

In this subsection, we analyse the treewidth and the planarity of the networks. The parameter values are shown in Table 5.13. The treewidth tends to be higher in inner cities, with Inn2 having the highest value. However, the gaps between inner cities, old town and residential areas are not that large. For the rural areas we observe widely spread values, and for the industrial areas quite low values.
The planarity shows that the inner cities and the old town are not planar, while both industrial areas are. Both rural areas and residential areas have both planar and non-planar instances.

5.5.2. Evaluation

The treewidth and the planarity do not seem to be that useful for classification since the results are quite heterogenous. However, it seems that the inner cities tend to have a higher treewidth and are not planar. It is also noteworthy that among the residential areas, Res2 again resembles the inner cities the most. Lastly, the old town seems to be more similar to the inner cities than to the residential areas again.

5.6. Overall Evaluation and Results

In this section, we bring together our partial evaluations and discuss them. To recap, our goal is to decide whether the classification of gas networks into the assumed class set is reasonable, or whether we have to categorize the gas networks differently, or whether a meaningful categorization is not possible at all. In Table 5.14 and Table 5.15, we have again listed the parameters that we found to be the most meaningful and useful to characterise the gas networks. In addition to these tables, for each parameter we also provide a number line on which the values of all instances are plotted. The colour and shape of the markers denote the class label of the instance.

Table 5.14.: Most useful parameters: length, inner diameter, diameter, node degree, average pressure.

Gas Network Instance	Parameter							
	Len_{max}	Len_{avr}	Len_{mdn}	ID_{avr}	$ID_{low_{avr}}$	$Diam_n$	Deg_{avr}	Pr_{avr}
Inn1	5499	131.5	86.7	141.6	168.7	12.78	2.383	2.525
Inn2	2272	99.1	67.3	177.9	170.3	6.82	2.453	1.185
Old1	528	69.7	49.2	119.9	119.3	9.97	2.422	0.127
Res1	520	77.2	55.2	111.4	107.1	14.59	2.194	0.270
Res2	730	85.7	52.6	118.4	120.9	10.92	2.260	0.299
Res3	1221	75.7	49.0	107.8	107.8	11.26	2.159	0.385
Rur1	3046	105.0	47.1	88.7	91.1	40.42	2.137	0.740
Rur2	5678	96.8	49.7	97.8	90.1	19.75	2.133	2.360
Rur3	2418	114.6	70.7	86.2	85.9	19.66	2.186	1.873
Ind1	1884	167.4	129.9	133.7	133.7	61.20	2.068	0.528
Ind2	1628	275.9	186.3	185.1	184.2	107.11	2.115	3.521

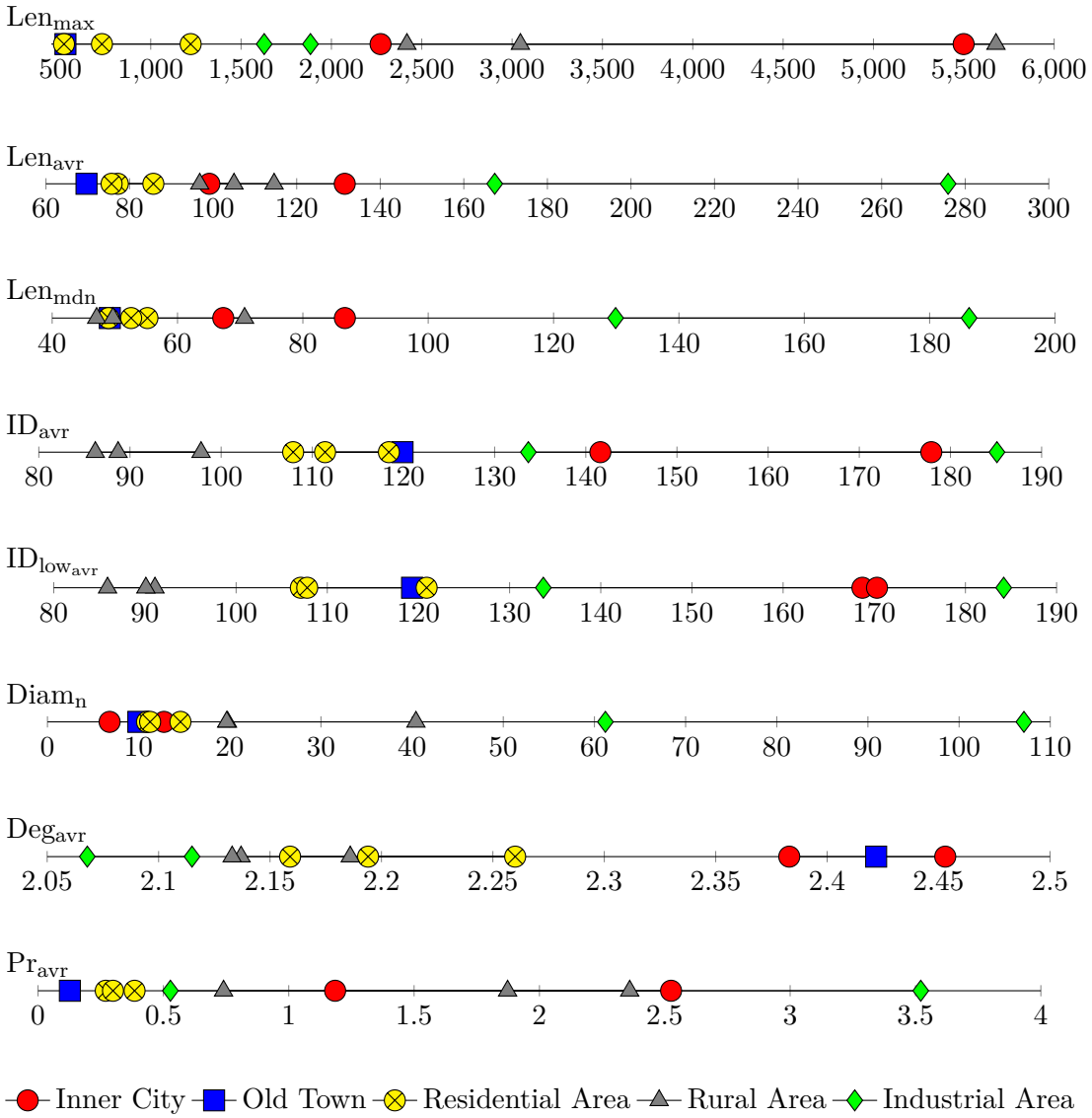
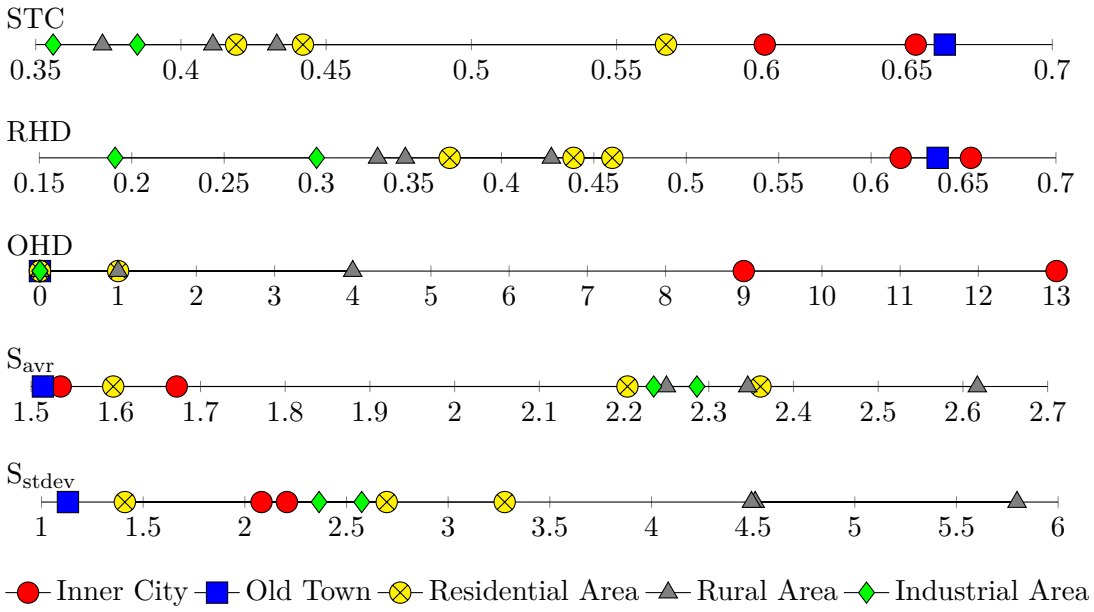


Table 5.15.: Most useful parameters: SizeTwocore (STC), RatioHigherDegree (RHD), OriginsHigherDegree (OHD), SizeTrees (S).

Gas Network Instance	**Parameter**				
	STC	RHD	OHD	S_{avr}	S_{stdev}
Inn1	0.601	0.616	9	1.672	2.082
Inn2	0.653	0.654	13	1.535	2.207
Old1	0.663	0.636	0	1.514	1.130
Res1	0.442	0.439	0	2.361	2.698
Res2	0.567	0.460	1	1.597	1.410
Res3	0.419	0.372	0	2.204	3.278
Rur1	0.411	0.333	0	2.346	4.510
Rur2	0.373	0.348	4	2.617	5.798
Rur3	0.433	0.427	1	2.250	4.492
Ind1	0.356	0.191	0	2.235	2.365
Ind2	0.385	0.300	0	2.286	2.575



In our evaluations we found that the parameter values seem to fit the assumed label set. While there are not many parameters where *each* class has a non-overlapping interval, there are many parameters where one or two classes have values that significantly differ from the other classes. For each class several parameters exist that are helpful to identify this class. As stated before, these parameters not necessarily separate those class instances from all other instances, but for example separate them together with another class from the others. However, we are then able to use another parameter to distinguish between these two classes. As an example, we look at the industrial areas and the inner cities: the parameter $ID_{low_{avr}}$ separates these two classes from the rest, and with the average node degree we can distinguish between them.

On the other hand, there were also many parameters where the values were very different both within each class and between the classes. These parameters are therefore not useful for classification, but they do not contradict our assumed label set.

It also often occurs that the instances of multiple classes have quite similar values. In general, this might indicate that these classes are too similar to each other and should

be combined to only one class. Still, we find that all classes have enough features or combinations of features that clearly distinguish them from the other classes. All classes even have at least one unique feature. By this we mean a parameter where the interval of the class does not overlap with the intervals of the other classes and ideally has a fairly wide gap to the other intervals.

For many parameters, the intervals of the residential areas are quite small, i.e., the values are very similar among the instances. For the rural areas, industrial areas and inner cities we observed larger intervals on many parameters. This indicates that residential area networks may have a more common structure with respect to these parameters than the other classes networks have.

We also found that the old town network stands somewhat out from the other classes as it seems to be a mixture of residential area and inner city characteristics. For most parameters, Old1 is very similar to either the residential areas or the inner cities and barely has values that do not fit into the intervals of either of them. We can roughly describe the old town with two main characteristics: first, regarding the lengths and extent parameters, it looks like the residential areas. Second, regarding the meshedness parameters, it is very similar to the inner cities.
Still, with respect to the combination of the parameters, the old town network is clearly different from both the residential areas and the inner cities. Therefore (based on Old1), we find old towns unique enough to be an own class.

Another very interesting observation was related to the residential area network Res2. While that network is very similar to the other two residential areas for many parameters, it is sometimes more similar to the inner cities than to the residential areas. For some parameters it is located between the inner cities and the other residential areas. While we need more data to further examine this observation, it indicates that the transition of values between residential areas and inner cities may be blurred with respect to some parameters. If this hypothesis can be confirmed, this can mean two things that have to be examined. On the one hand, it may mean that these parameters cannot be used for classification, or at least to a lesser extent. On the other hand, it can indicate that the two classes are too similar to each other so that they should only be one class. Of course, it is also well possible that Res2 is an outlier with regard to these parameters. For now, we still see enough indication that these classes can be distinguished quite well, but we state that this should be reviewed with more data.

Our overall results are that, based on our given data set, we can confirm the existence of the assumed gas network types. We found that the instances of each class have common features that we consider meaningful enough to assign these instances to the same class. On the other hand, the differences between the classes are significant enough to distinguish between them. We also found parameters that we find important and meaningful to classify gas network instances.
Nevertheless, we again state that these results are based on our data set and have to be reviewed when more data is available. This is especially true for the differences between inner cities and residential areas: we observed that Res2 is an instance that, for a subset of parameters, tends to be more similar to inner cities instances than to the other residential areas instances.

5.7. Analysis of the Supraregional Gas Network

In this section, we briefly compare the supraregional network with the other networks and classes of our data set. That network, denoted as Sup1 and shown in Figure 5.1, consists

Table 5.16.: Parameters used for the supraregional network analysis: length, inner diameter, diameter, node degree, average pressure.

Gas Network Instance	**Parameter**							
	Len_{max}	Len_{avr}	Len_{mdn}	ID_{avr}	$\text{ID}_{\text{low}_{\text{avr}}}$	Diam_{n}	Deg_{avr}	Pr_{avr}
Inn1	5499	131.5	86.7	141.6	168.7	12.78	2.383	2.525
Inn2	2272	99.1	67.3	177.9	170.3	6.82	2.453	1.185
Old1	528	69.7	49.2	119.9	119.3	9.97	2.422	0.127
Res1	520	77.2	55.2	111.4	107.1	14.59	2.194	0.270
Res2	730	85.7	52.6	118.4	120.9	10.92	2.260	0.299
Res3	1221	75.7	49.0	107.8	107.8	11.26	2.159	0.385
Rur1	3046	105.0	47.1	88.7	91.1	40.42	2.137	0.740
Rur2	5678	96.8	49.7	97.8	90.1	19.75	2.133	2.360
Rur3	2418	114.6	70.7	86.2	85.9	19.66	2.186	1.873
Ind1	1884	167.4	129.9	133.7	133.7	61.20	2.068	0.528
Ind2	1628	275.9	186.3	185.1	184.2	107.11	2.115	3.521
Sup1	14 266	338.0	1.7	164.4	148.4	56.54	2.139	32.010

of 3888 nodes and 4158 edges. As mentioned in the introduction, Sup1 is a different type of gas network than the regional networks we analysed so far. The purpose of networks of this type is to transfer large amounts of gas over great distances to supply the regional networks. Thus, supraregional networks cover larger areas and transfer the gas with much higher pressure. For these reasons, we do not consider supraregional networks a class of regional gas networks but rather a different type of network. Still, we think it is interesting to briefly look at the differences and common features between this network type and the regional networks. For these analyses, we focus on the parameters we found to be the most meaningful in the previous analyses (Table 5.16 and Table 5.17).

Starting with the pipeline lengths, we observe that the average and the maximum length of Sup1 is significantly higher than for all other networks. Interestingly, the median length is extremely small, i.e., the network seems to consist of both exceptional short and long pipelines.
For the average inner diameter of the pipelines, we observe no noticeable differences to the other networks, with the values of Sup1 ranking quite in the middle.
The normalised diameter of the supraregional network is smaller than those of the industrial areas, but noticeably larger compared to the other networks.
The average node degree of Sup1 is very similar to the rural area class, and in general does not stand out much from the other classes.
As expected, the average pressure is where the supraregional network differs heavily from all other instances. With a value slightly above 32 bar, the pressure of Sup1 is more than ten times greater than the maximum among the other instances.
Another parameter where Sup1 stands out from the other instances is the number of origin nodes with degree ≥ 3. It is significantly higher than the inner city instances, and consequently differs even more from the other classes.
For all remaining parameters (SizeTwocore, RatioHigherDegree, $\text{SizeTrees}_{\text{avr}}$, $\text{SizeTrees}_{\text{stdev}}$) we observe no noticeable differences between the supraregional network and the regional networks. For all of these parameters, Sup1 never has the minimum or maximum value among all instances but ranks somewhere between them.

Table 5.17.: Parameters used for the supraregional network analysis: SizeTwocore (STC), RatioHigherDegree (RHD), OriginsHigherDegree (OHD), SizeTrees (S).

Gas Network Instance	**Parameter**				
	STC	RHD	OHD	S_{avr}	S_{stdev}
Inn1	0.601	0.616	9	1.672	2.082
Inn2	0.653	0.654	13	1.535	2.207
Old1	0.663	0.636	0	1.514	1.130
Res1	0.442	0.439	0	2.361	2.698
Res2	0.567	0.460	1	1.597	1.410
Res3	0.419	0.372	0	2.204	3.278
Rur1	0.411	0.333	0	2.346	4.510
Rur2	0.373	0.348	4	2.617	5.798
Rur3	0.433	0.427	1	2.250	4.492
Ind1	0.356	0.191	0	2.235	2.365
Ind2	0.385	0.300	0	2.286	2.575
Sup1	0.431	0.322	21	1.947	3.850

Overall, with regard to most parameters the supraregional network is quite similar to the other classes. Its values are often particularly similar to the rural area and industrial area class. This comes to no surprise as these classes tend to be sparser and cover larger areas than the other classes, and we expected similar observations for the supraregional network. However, Sup1 stands out clearly from all other instances with regard to the maximum and average pipeline lengths and the average pressure. Again, these observations match our expectations, especially the significantly higher pressure. A very remarkable and somewhat surprising characteristic is the extremely small median length.

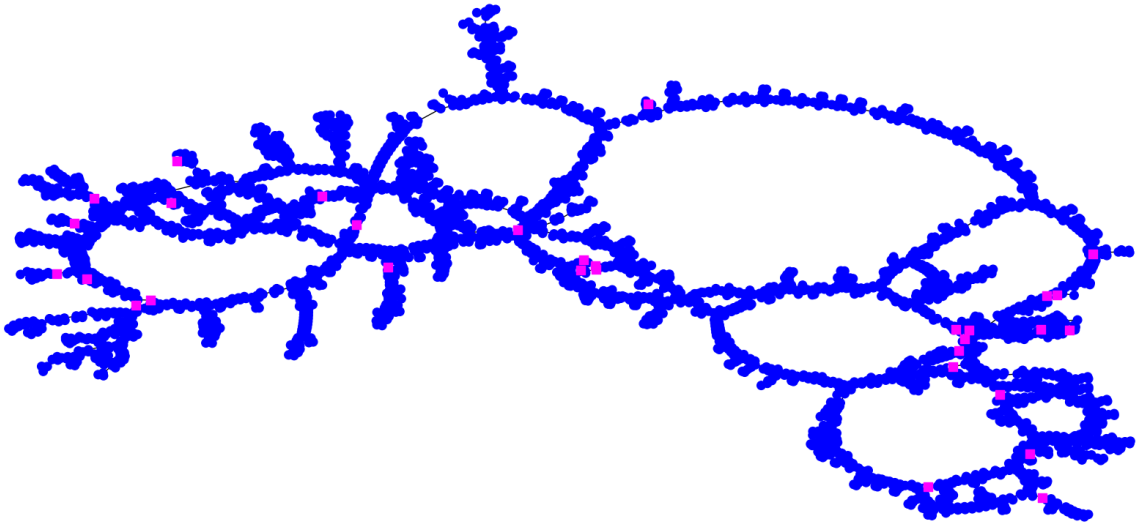


Figure 5.1.: Supraregional gas network instance Sup1.

6. Construction of three Gas Network Classifiers

In this chapter, we construct three gas network classifiers: the *Decision Tree Classifier* 6.1, the *Scoring System Classifier* 6.2 and the *Unique Feature Classifier* 6.3. While the decision tree is a well-known classifier, the other two approaches are developed by us. All three classifiers have in common that we use subsets of the parameters analysed in the previous chapters, but they are all based on different approaches. This means that they differ in the choice, number and usage of the parameters.
Two of the approaches use intervals, i.e., for each class and each parameter, the minimum and maximum parameter values are used to define an interval of values that are valid for that class. That means that for each parameter, the values of all instances of that class are within this interval. As we have only one old town network, these intervals contain only one point which is why we manually expand these intervals in two of the classifier approaches.

6.1. Decision Tree Classifier

Our first classifier is the *Decision Tree Classifier* (DTC) that we explained in Section 2.2.2. With the parameters that we figured out to be the most meaningful (see Table 5.14 and Table 5.15), there are several possibilities of constructing a decision tree based on our data set. To recap, the idea is to iteratively choose parameters that divide the remaining instances into two or more branches such that for each class all of its instances belong to the same branch. Since most of the possible decision trees use only a small number of parameters, we decide to construct three decision trees that use different subsets of the mentioned parameters. The parameters for each split are chosen manually, the thresholds are chosen to be robust to possible expansions of the class intervals that may occur with more data. The resulting decision trees are shown in Figure 6.1, Figure 6.2 and Figure 6.3. Note that these trees are only three examples of possible trees and there are several more combinations of parameters that can be used to construct decision trees.

The straightforward way of classifying gas network instances with these decision trees is to choose one of them, traverse it and assign the class label of the corresponding leaf. With more data and a test set available, the accuracies of the decision trees can be compared beforehand to determine the best decision tree that is then chosen as the final classifier.
Another approach is to combine all three (or even more) decision trees. To do so, an instance that is to be classified is assigned one class label from each of the decision trees.

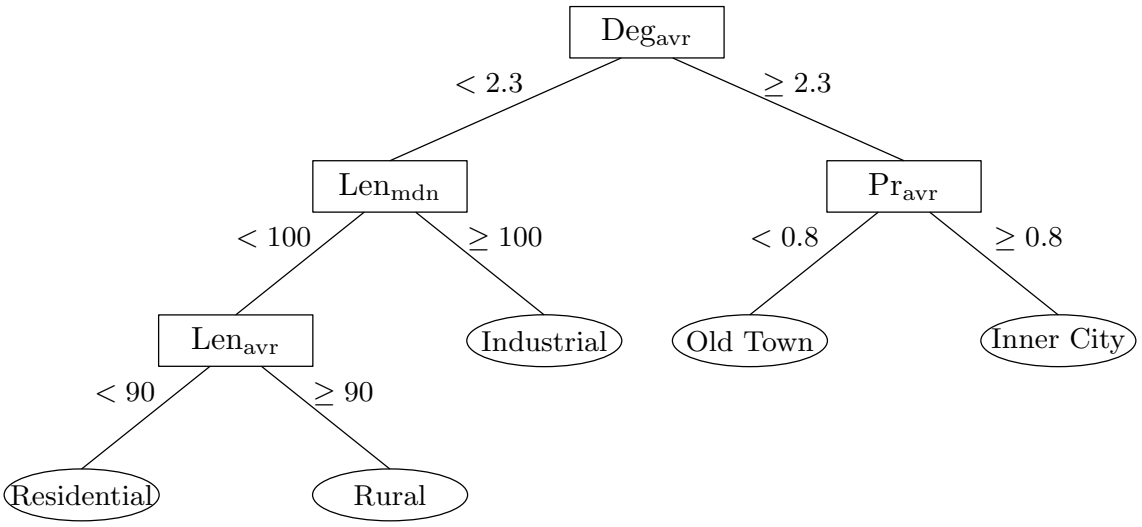


Figure 6.1.: First decision tree. Used parameters: Deg_{avr}, Len_{mdn}, Pr_{avr}, Len_{avr}.

The final label is then chosen via majority vote, i.e., the final class label of the instance is the one that was assigned to the instance by the most trees. This combination of different decision trees is similar to a technique known as bagging [Leo96]. The goal of this method is to reduce the variance: if the instance is an outlier with respect to one parameter, it will probably be classified wrongly by a decision tree that uses this parameter. In contrast, the other decision trees that do not use this parameter at all or use it in a different way will probably classify the instance correctly. With the majority vote the most probably correct label can then be assigned to the instance as the final label.

6.2. Scoring System Classifier

The second classifier is the *Scoring System Classifier* (SSC). The key feature of this approach is that it not only assigns a class label to an instance, but also computes likelihoods for each class that indicate how likely it is that the instance belongs to that class. This way we classify gas networks, but in addition have an indicator of how certain we are in that decision. This is especially interesting when more than one class have similarly high likelihoods or when an instance has quite low likelihoods for all classes.

6.2.1. Construction and Classification

In the first construction step, we select a set of parameters that we use for the classification step. Second, each of these parameters is assigned a weight indicating how meaningful we consider this parameter to be. The classification of an instance is then performed in the following way. First, we initialise the score of each class with zero. Next, we iterate over all selected parameters. For each parameter, we take the parameter value of the instance and check for each class if that value is within the class interval of the parameter. If it is, we increment the score of the class by the weight of the parameter. If it is not, the score is not increased. After iterating over all parameters, each class has a score which, in relation to the maximum possible score, indicates the likelihood that the instance belongs to that class. The label assigned to the instance is then the class with the highest likelihood. If two or more classes have the same likelihood, we define that the label *undefined* is assigned to the instance instead. Another possible way of handling these cases is to define specific tie-breaker rules.

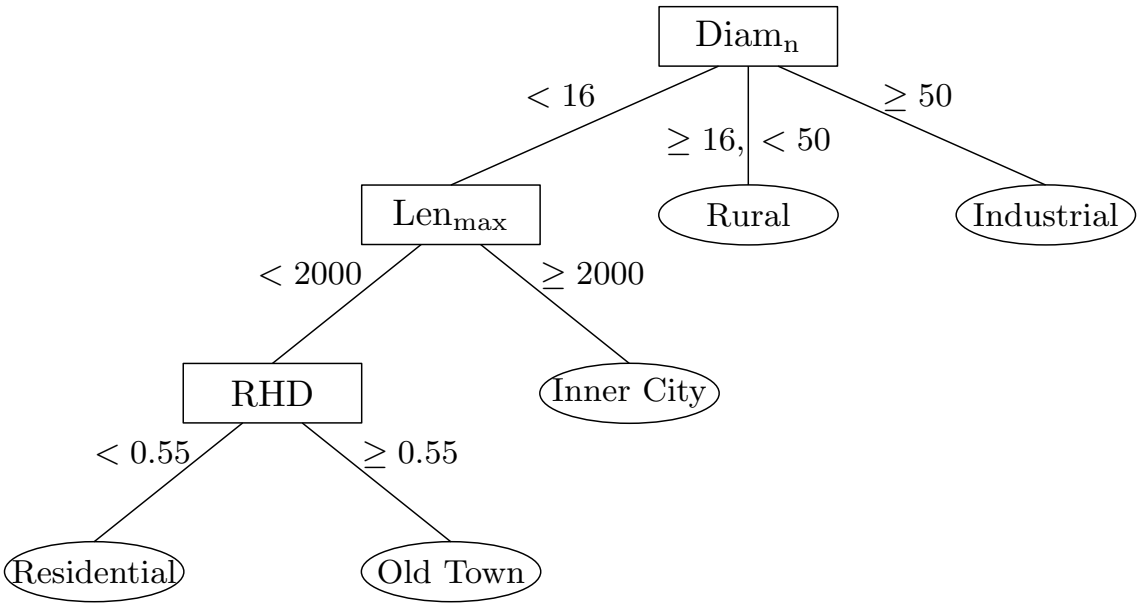


Figure 6.2.: Second decision tree. Used parameters: Diam_n, Len_{max}, RatioHigherDegree (RHD).

Table 6.1.: Scoring system classifier: selected parameters and weights.

Parameter	Weight
Len_{max}	1
Len_{avr}	2
$\text{ID}_{\text{low}_{avr}}$	2
Diam_n	2
Deg_{avr}	3
Pr_{avr}	3
SizeTwocore	1
OriginsHigherDegree	1
SizeTrees_{avr}	1
SizeTrees_{stdev}	1

As mentioned before, the intervals of the old town class consist of only one point. This results most probably in a score of zero for almost all instances since the instances will barely have an exact same parameter value as Old1. Therefore, we manually expand the intervals of the old town class for the selected parameters. For all but the average node degree we expand the intervals by 20 % in both directions. The average node degree is expanded by only 2 % in both directions, the reason being that this parameter differs between the classes only after the decimal point.

6.2.2. Selection of Parameters and Weights

We decide to select ten of the parameters we figured out to be the most meaningful (see Table 5.14 and Table 5.15). The selected parameters and their weights are shown in Table 6.1. In the following, we briefly explain our choices.

From the aforementioned most meaningful parameters we use all but the median length, the average inner diameter and the ratio of nodes with degree ≥ 3. The reason for this is that these parameters are quite redundant since we already use the average length, the average inner diameter of the low pressure graph and the size of the 2-core.

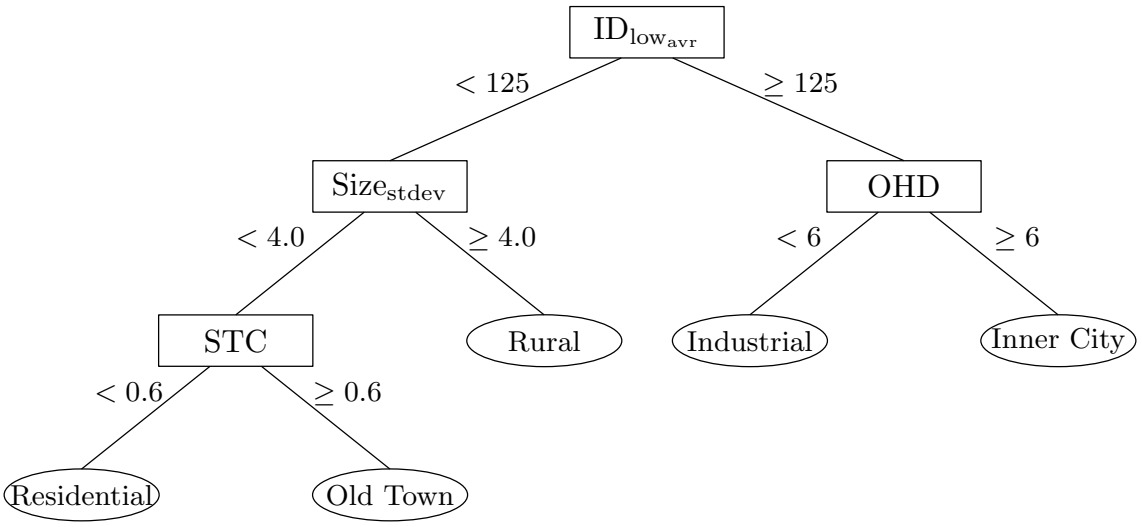


Figure 6.3.: Third decision tree. Used parameters: $ID_{low_{avr}}$, $SizeTrees_{stdev}(Size_{stdev})$, SizeTwocore (STC), OriginsHigherDegree (OHD).

For the weights, we consider the length, the average node degree and the average pressure the most important parameters. Thus, we assign them a weight of 3, where the length is further divided into the maximum (weight of 1) and the average length (weight of 2). A bit less, but still very important we consider the average inner diameter and the network diameter, which we each assign a weight of 2. While the other four parameters seem to be useful for classification, they are somewhat harder to interpret in the context of gas networks. Therefore, we consider them less important compared to the previous parameters and assign each of them a weight of 1. In total, these weights add up to 17 that is therefore the maximum possible score a class can get when classifying an instance.

6.2.3. Introduction to the Bayes Classifier

As explained before, the SSC is a probabilistic classification approach that we constructed specifically for our gas network classification problem and our given data set. One of the best known and widely used probabilistic classifiers is the *Bayes classifier* [Ber18] (BC). In this subsection, we briefly introduce that classifier and explain why we currently cannot use it for our data set. We do so because we think that with more data available, the BC is very suitable for our gas network classification problem.

Note that for reasons of simplicity, in the following we assume that all parameters are independent from each other. This assumption leads to the so-called *naive Bayes classifier* (NBC). Without this assumption the main idea stays the same, but details get more complicated, so we limit this explanation to the simple version.
The basic idea of the NBC is to compute the probability distribution for each parameter and class. With this distribution we can calculate the probability that the parameter value of a given instance appears in that class. If this probability is high, it indicates that (with regard to this parameter) the instance is more likely to belong to that class. These probabilities are calculated for all parameters and all classes. Then the probabilities of each class are put together into a single scalar, for example by multiplying all of them. These scalars are then the probabilities of the instance belonging to each class and can be compared among the classes.
The difficult part of constructing an NBC is to determine the probability distributions. With continuous parameter values, the Gaussian distribution (also known as normal distribution) is often found suitable. It is defined by the arithmetic mean and the standard deviation. For

discrete parameter values, the relative frequency of each value can be used as a probability distribution.

The availability of these probability distributions is the main problem that we face when we want to construct a BC for our data set. Since we have at most three samples available for each class, and for some classes only two or one, it is not possible to compute realistic probability distributions. With more data available, we think that for example Gaussian distributions can describe the distributions of many parameters quite well, and that the BC can provide good results, given good probability distributions. Another problem to be dealt with is the dependency of parameters, as we assume that not all parameters are independent from each other. One way to deal with this is to find a subset of independent parameters.

6.3. Unique Feature Classifier

Our third classifier is the *Unique Feature Classifier* (UFC). The idea of this approach is to find a unique feature for each class that characterises this class.

6.3.1. Construction and Classification

The construction of the UFC consists of selecting a unique feature for each class. By a unique feature of a class we mean a parameter where the corresponding class interval does not overlap with the intervals of all other classes. Ideally, it additionally has a fairly wide gap to the other intervals.
If the parameter value of an instance is within the interval of the unique feature of a class X, we say that the instance *fulfils* the unique feature of X. The interpretation of these unique features is the following: first, *all* instances belonging to X fulfil the unique feature of X. This property follows from the definition of the intervals. Second, *only* instances belonging to X fulfil the unique feature of X. This property follows from the definition of a unique feature since the unique feature interval of X is disjunct from the corresponding intervals of all other classes. Thus, when an instance fulfils the unique feature of a class, it belongs to that class.

Assuming the chosen features and intervals are indeed unique features and the interval boundaries are correct, each instance fulfils the unique feature of exactly one class. However, with outliers or with chosen features that are not unique, an instance may also fulfil the feature of no class or fulfil the features of multiple classes.

The classification of a given instance with a set of chosen unique features works as follows: for each class, we check how many unique feature a given instance fulfils. There are three possible outcomes: the instance fulfils exactly one unique feature, the instance fulfils no unique feature or the instance fulfils more than one unique feature. In the first case, the classifier assigns the instance to the class whose unique feature is fulfilled. In the second and the third case, the interpretation is that the instance belongs to either no class or to multiple classes, both of which is not intended. We define that the classifier assigns the label *undefined* in these cases. As for the SSC, instead it is also possible to define specific tie-breaker rules for these situations.

6.3.2. Selection of Unique Features

We now propose two parameters for each class that can be used as unique features, i.e., they satisfy the definition of a unique feature. As explained earlier, the intervals of these features follow directly from the minimum and maximum values in the instances of the class (see Table 5.14 and Table 5.15 and the corresponding number lines).

For the industrial area class, we suggest the median of the pipeline length and the normalised network diameter. Both of these parameters have significantly lower values for all other classes and are therefore well suited to characterise industrial networks. Formally, the values for an industrial area instance have to be within the following intervals:

$$\mathrm{Len}_{\mathrm{mdn}} \in [129.9, 186.3]\,, \tag{6.1}$$

$$\mathrm{Diam}_{\mathrm{n}} \in [61.20, 107.11]\,. \tag{6.2}$$

The normalised diameter is also a possible unique feature for the rural areas. While the interval is neither at the lower nor at the upper end of the different class intervals, it has fairly large gaps to the residential areas at the lower boundary and to the industrial areas at the upper boundary. As the second possible unique feature we choose the average inner diameter. For this parameter the rural area instances have the lowest values compared to all other classes. The difference to the other classes is even more significant on the low pressure graph, which is why we take the average inner diameter on this graph instead of the complete graph. The resulting intervals of these unique features of rural areas are:

$$\mathrm{Diam}_{\mathrm{n}} \in [19.66, 40.42]\,, \tag{6.3}$$

$$\mathrm{ID}_{\mathrm{low}_{\mathrm{avr}}} \in [85.9, 91.1]\,. \tag{6.4}$$

The parameters we propose for the residential area class are the average pipeline length and the average pressure. In comparison to residential areas, both of these parameters have smaller values in old towns and greater values in all other classes. The intervals of these features are as follows:

$$\mathrm{Pr}_{\mathrm{avr}} \in [0.270, 0.385]\,, \tag{6.5}$$

$$\mathrm{Len}_{\mathrm{avr}} \in [75.7, 85.7]\,. \tag{6.6}$$

For the inner cities, we suggest the number of origin nodes with degree ≥ 3 and the relative size of the 2-core. For the former, the inner cities have significantly larger values than all other classes. For the latter, the values of the inner cities are located between the old town class and the other classes. The intervals of these parameters are:

$$\mathrm{OriginsHigherDegree} \in [9, 13]\,, \tag{6.7}$$

$$\mathrm{SizeTwocore} \in [0.601, 0.653]\,. \tag{6.8}$$

For the old town class, the first possible unique feature is the average pressure that is lower for the old town than for all other classes. As second unique feature we propose the standard deviation of the size of the attached trees. Again, the value for this parameter is the lowest for the old town class among all classes.
As seen before, for the old town class the intervals of all parameters, including the chosen unique features, consist of only one point. Thus, we manually expand these intervals to the following ones:

$$\mathrm{Pr}_{\mathrm{avr}} \in [0.100, 0.200]\,, \tag{6.9}$$

$$\mathrm{SizeTrees}_{\mathrm{stdev}} \in [1.000, 1.300]\,. \tag{6.10}$$

7. Evaluation and Discussion of the Classifiers

In this chapter, we evaluate the classification results of our three classifiers constructed in the previous chapter. We then discuss the strengths and weaknesses of the classifier approaches and compare them with each other. At the end of the chapter, we provide an outlook on how each classifier can be adjusted and refined given more data.

7.1. Evaluation

To evaluate the classifiers, we use each of them to classify our training set instances as we have no distinct test set available. Remember that we used that training set to construct the classifiers; that means, it determined the interval boundaries, the splits in the decision trees and the unique features. Thus, on exactly that data set, the accuracy of all classifiers is 100 %. Of course, we cannot overrate these results, as we stated earlier that using the training set to determine the accuracy overestimates the quality of a classifier.

The evaluation of the DTC does not provide more information than the classification results, i.e., the assigned class label. The same applies to the UFC, where each instance fulfils the unique feature of exactly one class.

The evaluation of the SSC provides some more information, as the results are not only the assigned class labels, but actual scores for each class. These scores are shown in Table 7.1. We see that each instance has the maximum possible score of 17 for the class it belongs to. More interesting are the scores they have for the other classes. For both inner city instances, we observe similarities to the old town and the industrial area classes as these have a score of 5. The scores for the rural area and residential area class differ between both instances. For the old town instance we see what we analysed before, namely some similarities to both the inner city and the residential area class. However, both scores are still fairly low, so we can state that we can clearly identify Old1 as an old town instance. Among all instances, the residential area instances have the highest score for another class, with scores of 6 and 7 for the old town class. That is not too surprising, as we already pointed out the similarities between these classes. Again though, the instances are still clearly identified as residential areas. Somewhat interesting is that Res2 only has a score of 3 for the inner city class, despite our observations that this instance is often quite similar to the inner cities. The rural area instances have quite low scores for the other classes in

Table 7.1.: Scoring system classifier: resulting scores of all instances and classes.

Gas Network Instance	**Class score**				
	InnerCity	OldTown	Residential	Rural	Industrial
Inn1	17	5	4	1	5
Inn2	17	5	1	5	5
Old1	5	17	4	1	1
Res1	0	6	17	2	1
Res2	3	6	17	1	0
Res3	2	7	17	5	1
Rur1	3	1	2	17	4
Rur2	3	0	0	17	4
Rur3	6	0	6	17	4
Ind1	0	1	3	2	17
Ind2	0	1	3	3	17

general, with Rur3 being an exception as it has a score of 6 for both the inner city and the residential area class. Also, all instances show some similarities to the industrial area class. Lastly, the industrial area instances have a maximum score of 3 for any other class, thus they have only very small similarities with the other classes.
In total, we can state that the SSC achieves fairly clear results, as the maximum score an instance has for a class it does not belong to is 7.

7.2. Discussion and Comparison

In this section, we discuss the strengths and weaknesses of each classifier and compare the different approaches. As all classifiers have an accuracy of 100 % on our data set, a comparison based only on the classification results is not helpful at all. Therefore, we discuss the classifiers not only with regard to our data set, but more generally.

Starting with the UFC, we think that this approach is quite simple, but accurate. If a unique feature for each class exists and its intervals is known, the classifier is very fast and simple in both construction and classification. However, the approach has some weaknesses, mainly related to the information needed for construction. The intervals of each unique feature have to be very precise, as an instance whose parameter value does not match the interval of the unique feature of a class is not assigned that class label, no matter how much the value deviates from the interval. Because of this, the approach is also very sensitive to instances where one specific parameter has an outlier value. Such values can also lead to fulfilled unique features of more than one class. Furthermore, it is not guaranteed that unique features exist for all classes.

Going on with the DTC, one advantage over the UFC is that we are not dependent on parameters that clearly separate one class from all others. While these are still helpful, we can also use parameters that separate the classes into two branches where multiple classes are on one branch. As the UFC, the DTC is also very sensitive to outlier values, as taking one “false” branch probably leads to a false classification result. An advantage of the DTC is that this weakness can be faced by using multiple decision trees and doing a majority vote. On the other hand, the multiple decision tree approach is likely to require a larger number of meaningful parameters than the UFC. Another strength of the DTC is that the interval boundaries of the used parameters do not have to be known exactly beforehand, as the threshold for the split of a parameter can be, for example, between the upper bound

of the interval of one class and the lower bound of the interval of another class. This way, both intervals are expanded and an instance that narrowly is not within an interval is still assigned the correct branch. For the UFC, the result for that parameter would be "not fulfilled".

Lastly, we discuss the SSC. Its biggest advantage over the other classifiers is that the classification result not only provides the assigned class label, but likelihoods for all classes. This way, ambiguous classification results are identified. Another advantage is that this approach is robust against outliers, as only the specific outlier values increase the scores wrongly, while all other parameters increase the scores as intended. Thus, few outlier values do not necessarily lead to wrong classification results. The probably biggest weakness of the SSC are the hard interval boundaries, as an instance that narrowly misses an interval of a class does not increase that class score at all. We observe this for the instance Res2: as seen in the analyses and in the number lines (Table 5.14, Table 5.15), Res2 is sometimes more similar to the inner city class than to the residential area class. But, as its parameter values are most often still below the intervals of inner cities, the score for the inner city class is still quite low (see Table 7.1).

Another point worth discussing is the number of meaningful parameters needed by each approach, as the availability of such parameters may differ from data set to data set. The UFC uses exactly one parameter per class, and it may even be possible to use a single parameter for multiple classes. The DTC uses only few parameters for one decision tree, but needs more for the approach with multiple decision trees. In general, the classification results of the DTC are likely to be more accurate the more parameters are used. The same is true for the SSC: while we decided to choose ten parameters, a smaller number will work too, but the accuracy most likely increases when using more parameters.

Summarised, we observed that all three approaches have their advantages and disadvantages. The quality of each classifier is also dependent on the number of meaningful parameters and thus may differ between different data sets. When the intervals are well-known upfront and unique features for all classes exist, the UFC is a simple, but precise classifier that uses a fixed and fairly small number of parameters. The DTC is less dependent on interval boundaries and is able to handle outlier values by combining multiple trees, but probably needs quite many parameters for the latter. Overall, we think that the SSC is the best classification approach, as it provides more detailed results and is very robust against outliers. This is especially true when many meaningful parameters are available. However, that approach is quite dependent on accurate interval boundaries.

7.3. Outlook: Adjusting the Classifiers with more Data

In this section, we give a brief outlook on how the different classifiers have to be adjusted and can be improved with more data.
In general and independent from the used classifier, the parameter intervals of each class have to be adjusted with the new data. Furthermore, we made the assumption that the valid parameter values for each class are indeed intervals, i.e., all values between the lower and upper boundary are valid. In fact, it may also be possible that this assumption is not true for all parameters. It may occur that the valid values are sets of intervals, with the values between these intervals not being valid. It is also possible that no intervals, but only a set of discrete values are valid for some parameters. Thus, our interval assumption has to be verified.

Looking at the DTC, it is possible that a currently used parameter does not split the classes the same way it does right now, as new overlaps between the class intervals can

occur. If even all class intervals overlap for the parameter, that parameter may not be useful at all anymore. In these cases, either additional splits are required, or the parameter has to be replaced by another one.

The adjustment of the SSC is the easiest one among the classifiers. As the approach does not need distinct intervals, new overlaps between class intervals do not entail any necessary changes. Still, if a parameter seems to be not meaningful anymore, it should not be used further. On the other hand, if a currently unused parameter shows to be meaningful, it can be added to the classifier.
As mentioned in the previous chapter, a similar but more powerful approach is the Bayes classifier. With enough data available, probability distributions can be calculated for the parameters to construct that classifier. Note that it is important to consider dependencies between the used parameters.

The UFC is probably the most vulnerable classifier with respect to new intervals. For the currently used unique features it has to be validated if they are still unique features, i.e., if their intervals still have no overlap with the intervals of all other classes. If this is not true for a unique feature, a new one has to be found. However, this may not be possible for all classes, as the existence of unique features depends on the data. In the case that for one or even more classes no unique feature exists, that classifier approach may not be applicable anymore.

8. Conclusion

In this thesis, we examined whether it is possible to classify gas networks into different classes based on the region type which an instance originates from. For this purpose, we modelled gas networks as graphs and examined them with a graph-theoretic approach.

In Chapter 4, we selected and developed several graph parameters to characterise graphs and gas networks from various perspectives, including both more general graph parameters and gas network specific parameters.
In the next step, in Chapter 5, we analysed a data set of labelled gas network instances, originating from five different regions. We computed the parameter values of these networks and pointed out differences and common features between the instances of each class and between the different classes. We found that the gas networks originating from different regions indeed show significant differences among each other while networks originating from the same class show significant common features. Although this is not true for all parameters, we found that each of the five regions has several characteristics that distinguish it from all other regions. Therefore, we concluded that a classification of gas networks into five distinct classes is reasonable. Furthermore, we pointed out the most meaningful parameters that are helpful to describe the different classes and to distinguish between them.
Regarding the old town class, we made the interesting observation that for most parameters, it is a mixture of the characteristics of the inner city class and the residential area class. Furthermore, we observed that one residential area instance tends to be more similar to the inner city class than to the residential area class regarding some parameters. Thus, we consider it possible that the borders between these classes are blurred, at least with respect to these parameters.

With the classes and meaningful parameters found, in Chapter 6 we then presented three different classifier approaches that assign class labels to gas network instances. In Chapter 7, we evaluated and discussed these approaches. We pointed out their advantages and disadvantages, and found that their applicability and quality depends on the data set. Overall, we found the Scoring System Classifier to be the most promising, as for a gas network it provides not only the final class label, but also likelihoods for the membership of that instance to each class.
Furthermore, we provided an outlook on the further development of each classifier approach. We explained how each of them has to be adjusted and how it can be improved with more data available.

8.1. Outlook

As mentioned above, we used a graph-theoretic approach to examine and classify gas networks. A completely different approach may be to use machine learning algorithms [KZP+07]. Although there exist no such specific algorithms to classify gas networks, classification and clustering are well-examined topics in the field of machine learning. Thus, it may be possible to model the gas networks in a way that known algorithms are applicable. However, both supervised and unsupervised machine learning algorithms rely on a large amount of data that may not be available in the future.

On the other hand, also our graph-theoretic approach can be developed further. As we already stated, all our findings are based on our specific data set. Once more data is available, these findings have to be reviewed to either confirm the classes we found or to adjust the set of classes. Also, the set of meaningful parameters may differ from the one we found. Then, the classifiers can be improved as we described in Section 7.3. Especially for the Scoring System Classifier we find it very promising to transform it into a Bayes classifier.

Another large field of further research is the generation of generic gas networks. Such generic networks should capture the most important characteristics of the class they belong to. As explained in the introduction, these instances can then be used for various objectives. One use case is to examine a generic network representing a particular region instead of the actual gas network of this region, for example if a model of the actual network is not available. Generic networks also enable simulations of the future evolution of regions, for example when a small city is expected to massively grow within the next years or when a significant decline in gas demand in a region is expected.
The parameters and the characteristics of each class found in this work may be helpful to generate generic instances of any adjustable size and region type. To choose the value of a parameter for a generic network, one approach may be to compute a randomised value that is within the valid class interval of that parameter. If a probability distribution of the parameter is available, a more accurate approach may be to draw the parameter value from that probability distribution rather than assuming a uniform distribution on the class interval. The generic network can then be constructed in a way that the parameter takes this parameter value, possibly with a small deviation.
We consider the 2-core and the attached trees to provide a good foundation for a generic gas network, as they describe the basic structure of the graph. The average node degree can be used to determine the ratio of nodes and edges the network should have, and parameters like the lengths and inner diameter describe how the pipelines should look like. Next, the parameters regarding the pressure stages and pressure areas can be used to determine the number and positioning of regulators.
Overall, we expect the generation of generic gas networks to be challenging, but also very promising and therefore an interesting field of further research.

Bibliography

[ACP87] Stefan Arnborg, Derek G. Corneil, and Andrzej Proskurowski. Complexity of Finding Embeddings in a k-Tree. *SIAM Journal on Algebraic Discrete Methods*, 8(2):277–284, 1987.

[Agg18] Charu C Aggarwal. An introduction to cluster analysis. In *Data Clustering*, pages 1–28. Chapman and Hall/CRC, 2018.

[Ber18] Daniel Berrar. Bayes' theorem and naive Bayes classifier. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 403, 2018.

[Dou12] Geoff Dougherty. *Pattern Recognition and Classification: An Introduction*. Springer Science & Business Media, 2012.

[DRC] The DVGW Research Center at the Engler-Bunte-Institute of Karlsruhe Institute of Technology. `https://www.dvgw-ebi.de/en/`. Accessed: 2022-09-11.

[ESA+21] Absalom Ezugwu, Amit Shukla, Moyinoluwa Agbaje, Adán José-García, Oyelade Olaide, and Ovre Agushaka. Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature. *Neural Computing and Applications*, 33, 06 2021.

[GEU] Key facts about gas in the EU. `https://acer.europa.eu/en/Gas/Documents/ACER_FACT-SHEETS_2021-07_02.pdf`. Accessed: 2022-09-11.

[KKU+17] Elisabeth Krueger, Christopher Klinkhamer, Christian Urich, Xianyuan Zhan, and P Suresh C Rao. Generic patterns in the evolution of urban water networks: Evidence from a large Asian city. *Physical Review E*, 95(3):032312, 2017.

[KZP+07] Sotiris B Kotsiantis, Ioannis Zaharakis, P Pintelas, et al. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24, 2007.

[LB17] Jingyi Lin and Yifang Ban. Comparative analysis on topological structures of urban street networks. *ISPRS International Journal of Geo-Information*, 6(10):295, 2017.

[Leo96] Breiman Leo. Bagging Predictors in Machine Learning. 1996.

[LWR16] Yizheng Liao, Yang Weng, and Ram Rajagopal. Urban distribution grid topology reconstruction via Lasso. In *2016 IEEE Power and Energy Society General Meeting (PESGM)*, pages 1–5. IEEE, 2016.

[LZWT14] Yangguang Liu, Yangming Zhou, Shiting Wen, and Chaogang Tang. A strategy on selecting performance metrics for classifier evaluation. *International Journal of Mobile Computing and Multimedia Communications (IJMCMC)*, 6(4):20–35, 2014.

[Min89] John Mingers. An empirical comparison of pruning methods for decision tree induction. *Machine learning*, 4(2):227–243, 1989.

[RFRW22] Manou Rosenberg, Tim French, Mark Reynolds, and Lyndon While. Finding an optimised infrastructure for electricity distribution networks in rural areas - A comparison of different approaches. *Swarm and Evolutionary Computation*, 68:101018, 2022.

[Sha19] Ayyoob Sharifi. Resilient urban forms: A review of literature on streets and street networks. *Building and Environment*, 147:171–187, 2019.

[Tan20] Suryakanthi Tangirala. Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, 11(2):612–619, 2020.

[TSB+20] Daniel Then, Christian Spalthoff, Johannes Bauer, Tanja M Kneiske, and Martin Braun. Impact of Natural Gas Distribution Network Structure and Operator Strategies on Grid Economy in Face of Decreasing Demand. *Energies*, 13(3):664, 2020.

[WM97] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

[YLLL22] Heng Ye, Zhiping Li, Guangyue Li, and Yiran Liu. Topology Analysis of Natural Gas Pipeline Networks Based on Complex Network Theory. *Energies*, 15(11):3864, 2022.

[ZSG19] Xia Zhu, Weidong Song, and Lin Gao. Topological characteristics and vulnerability analysis of rural traffic network. *Journal of Sensors*, 2019, 2019.

Appendix

A. Gas Network Data Set

A.1. Gas Network Instances

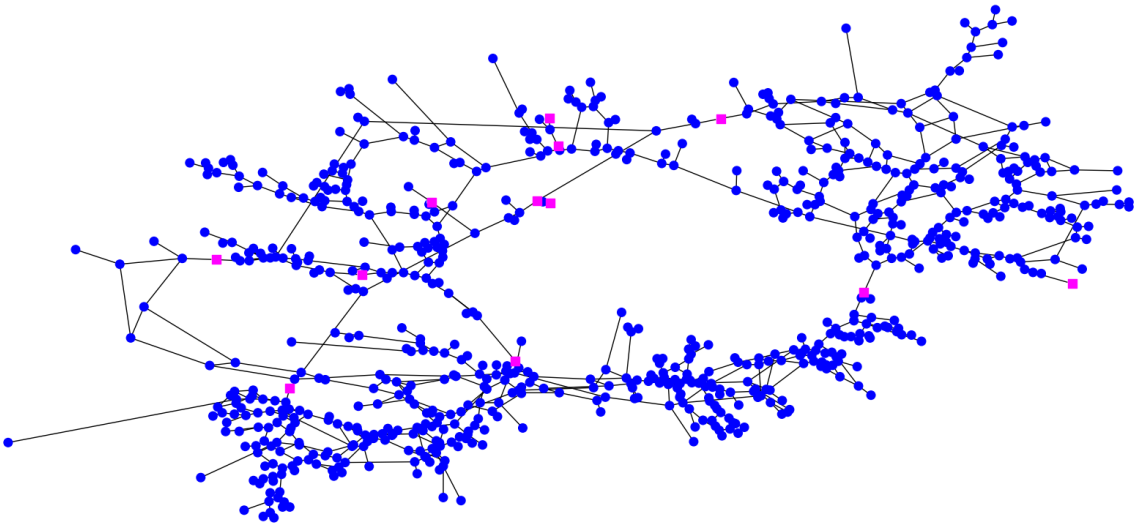


Figure A.1.: Gas network instance Inn1.

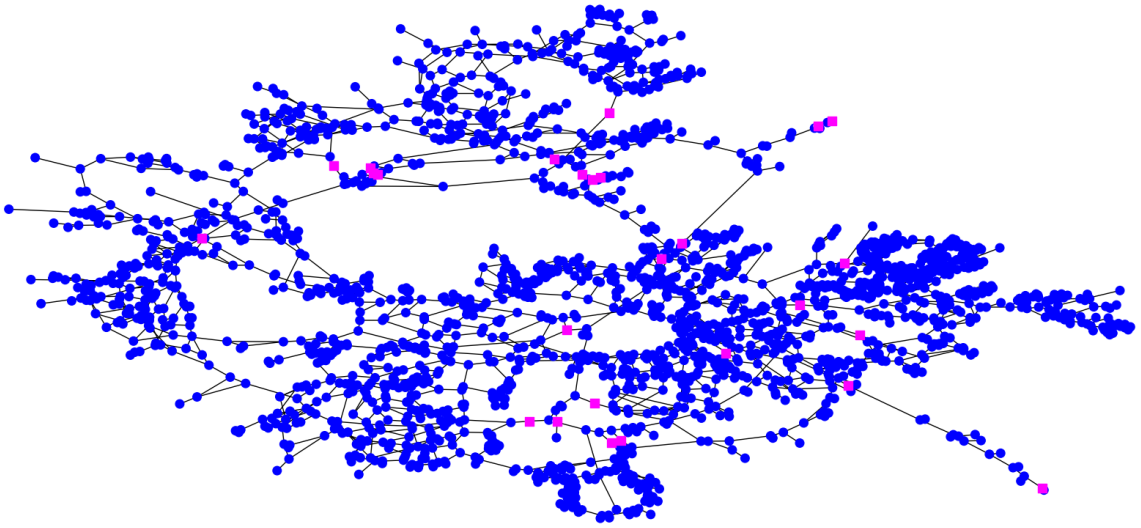


Figure A.2.: Gas network instance Inn2.

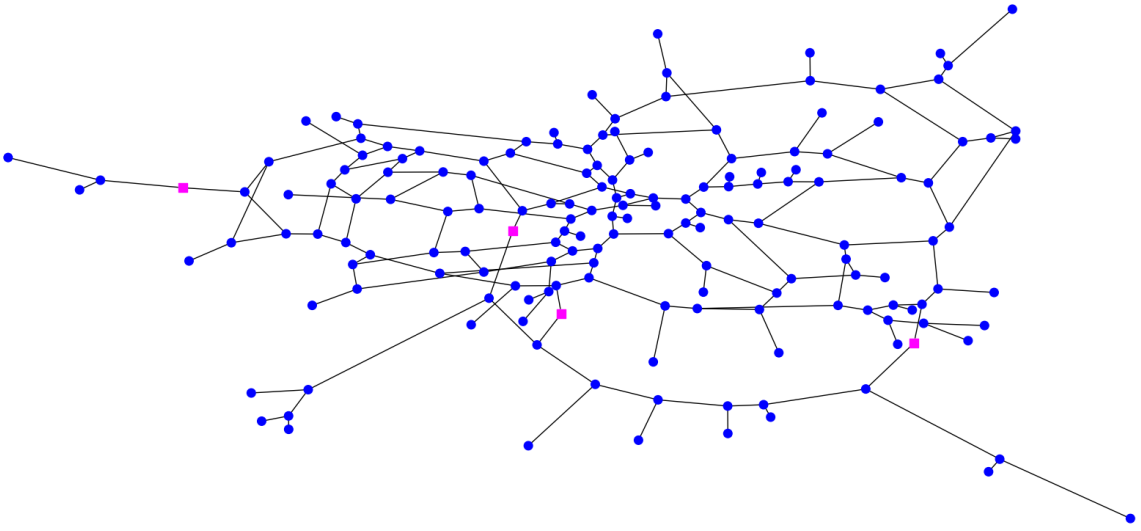


Figure A.3.: Gas network instance Old1.

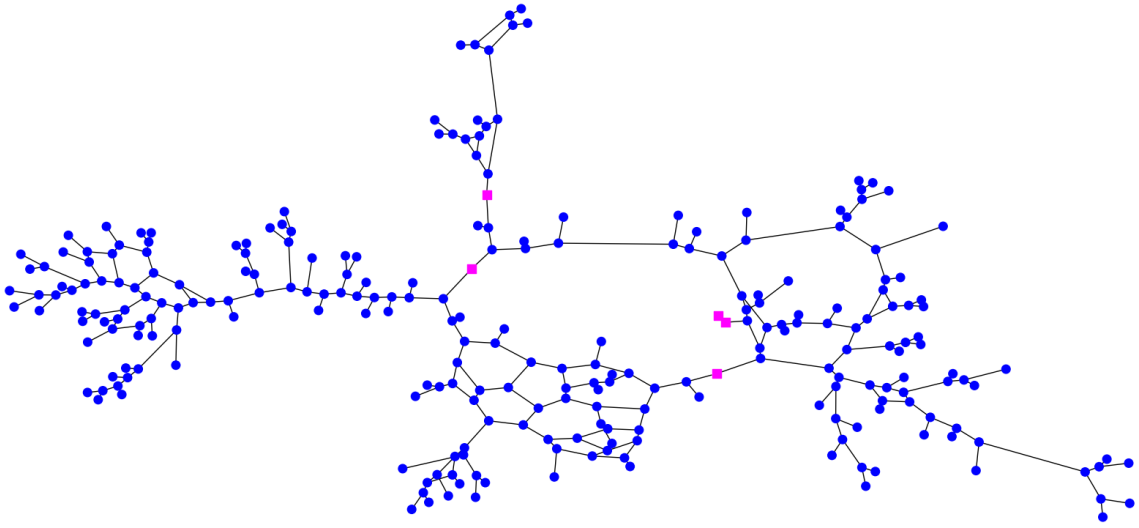


Figure A.4.: Gas network instance Res1.

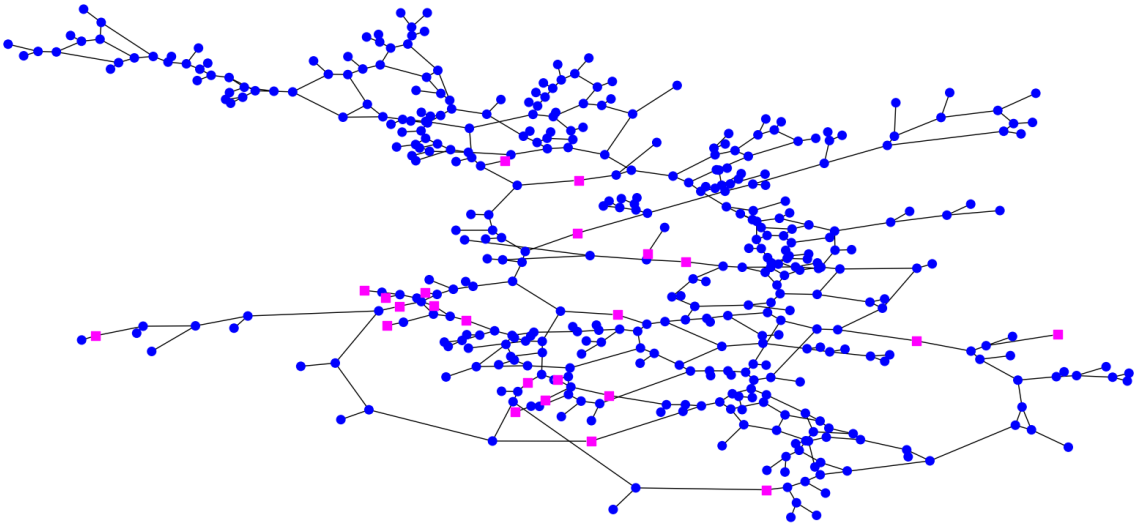


Figure A.5.: Gas network instance Res2.

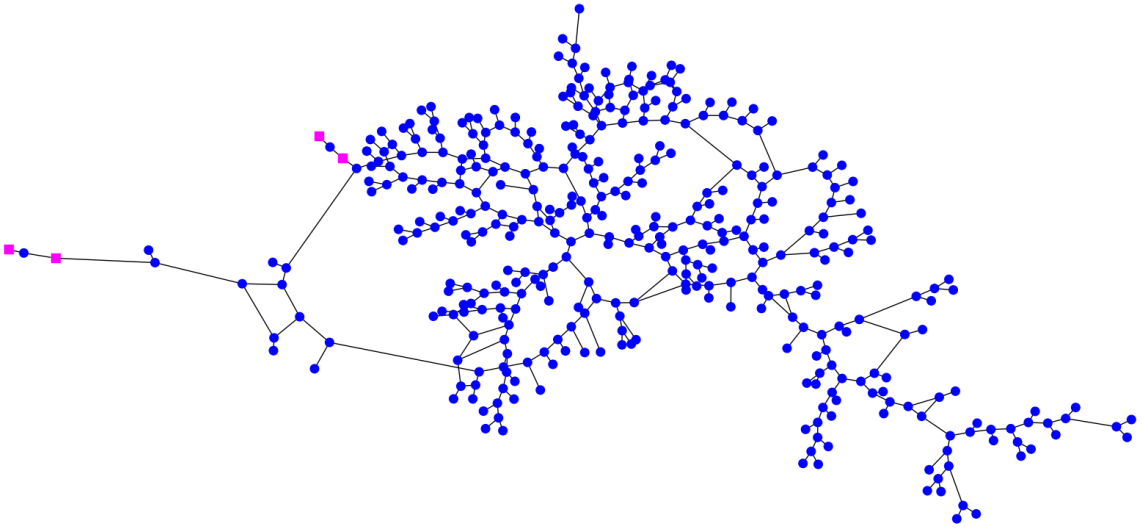


Figure A.6.: Gas network instance Res3.

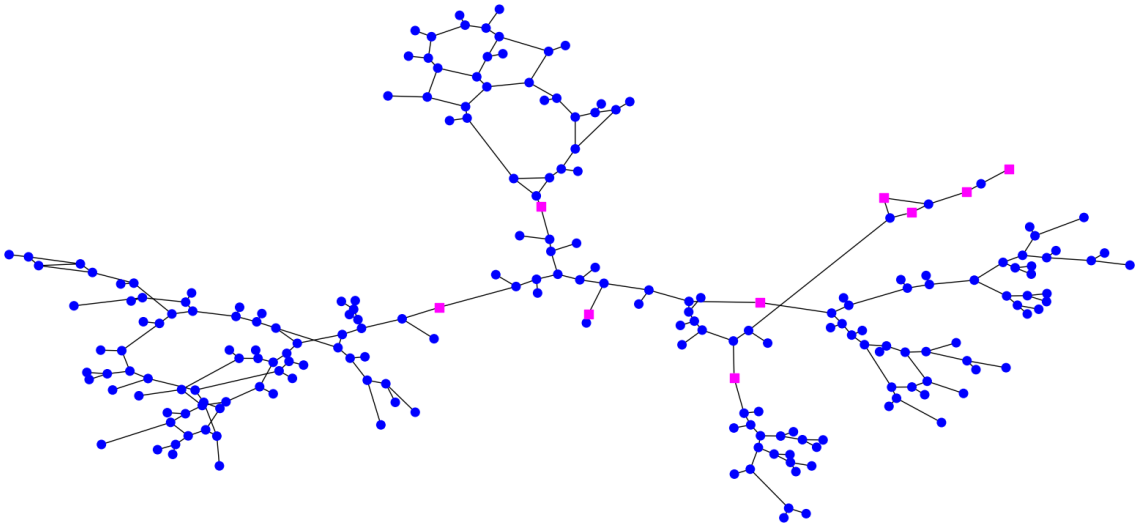


Figure A.7.: Gas network instance Rur1.

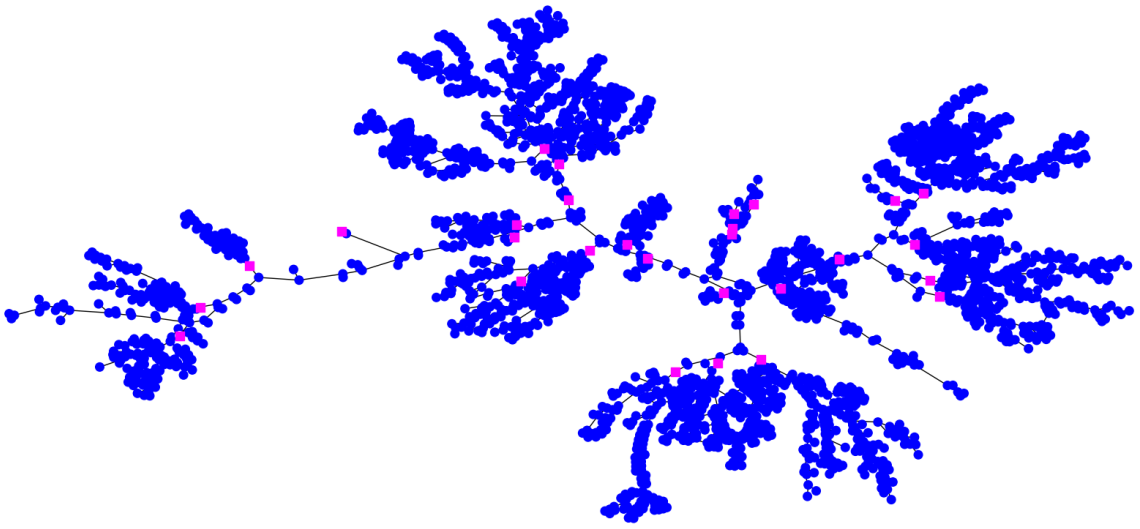


Figure A.8.: Gas network instance Rur2.

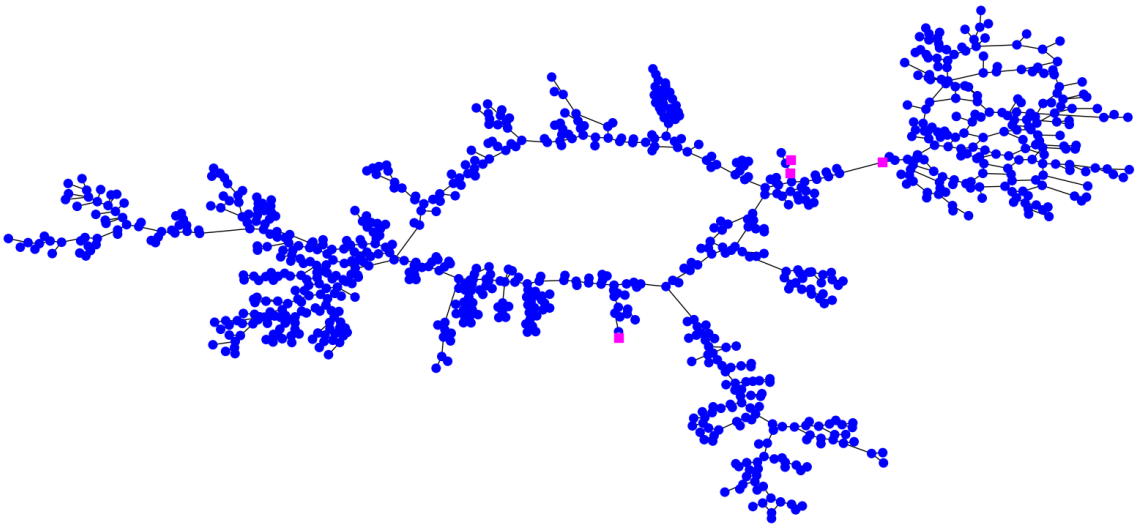


Figure A.9.: Gas network instance Rur3.

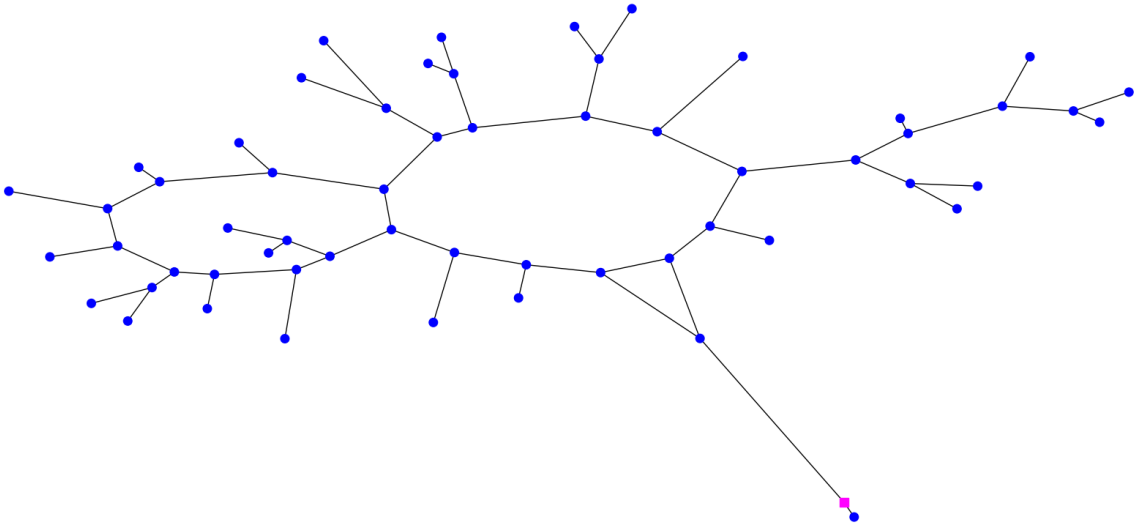


Figure A.10.: Gas network instance Ind1.

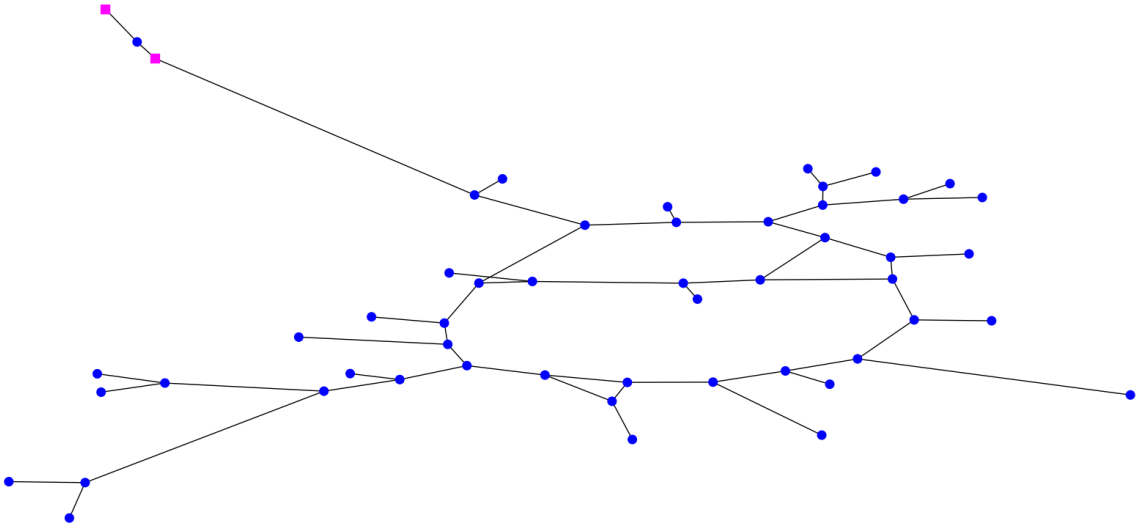


Figure A.11.: Gas network instance Ind2.

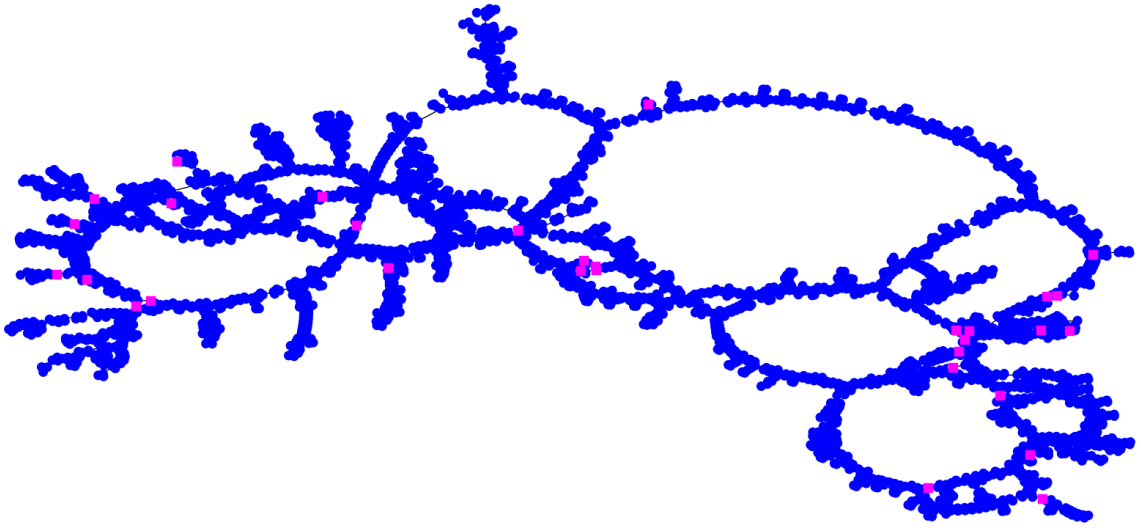


Figure A.12.: Gas network instance Sup1.

A.2. Networks with Pressure Stages

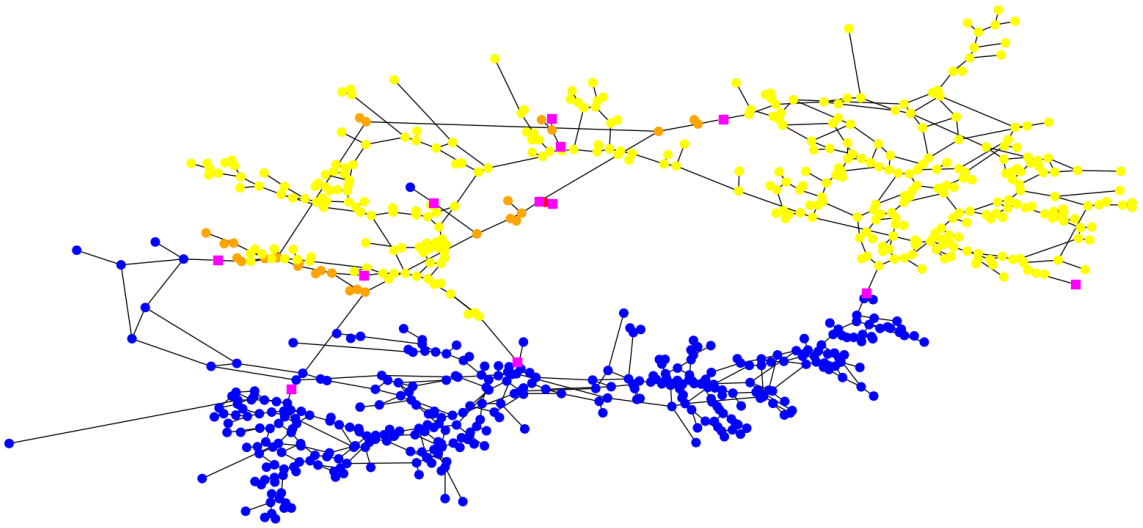


Figure A.13.: Pressure stages of instance Inn1.

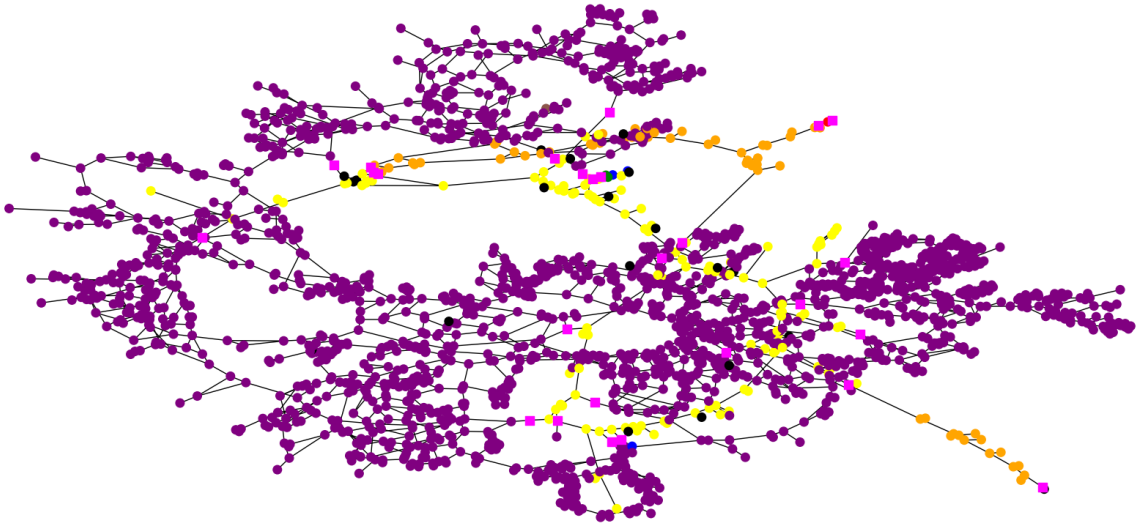


Figure A.14.: Pressure stages of instance Inn2.

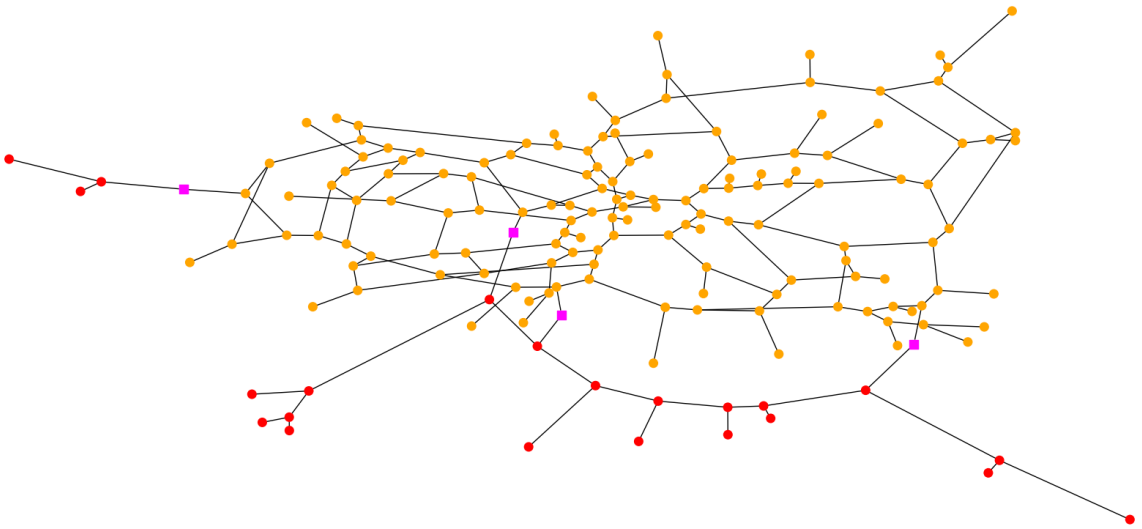


Figure A.15.: Pressure stages of instance Old1.

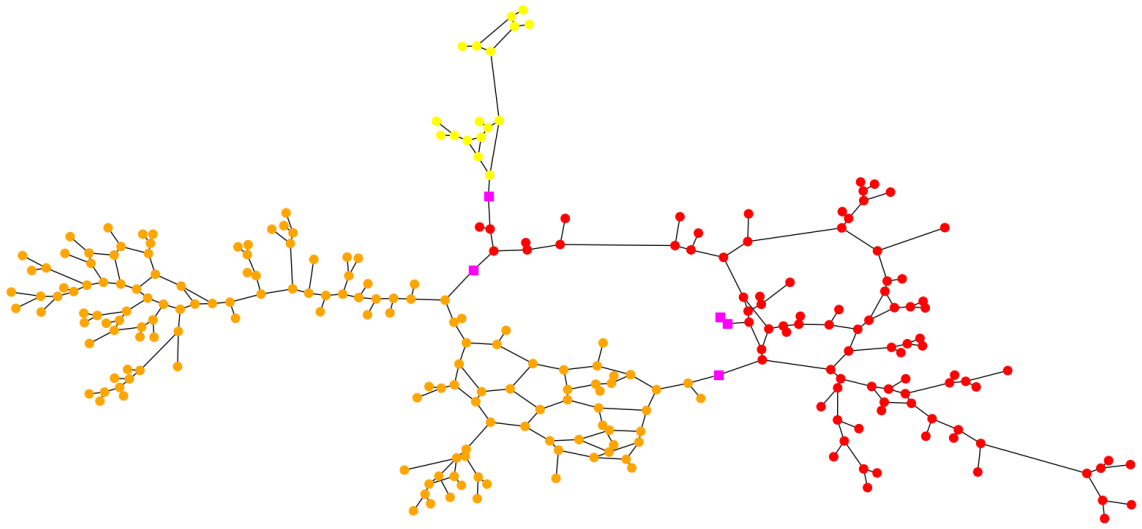


Figure A.16.: Pressure stages of instance Res1.

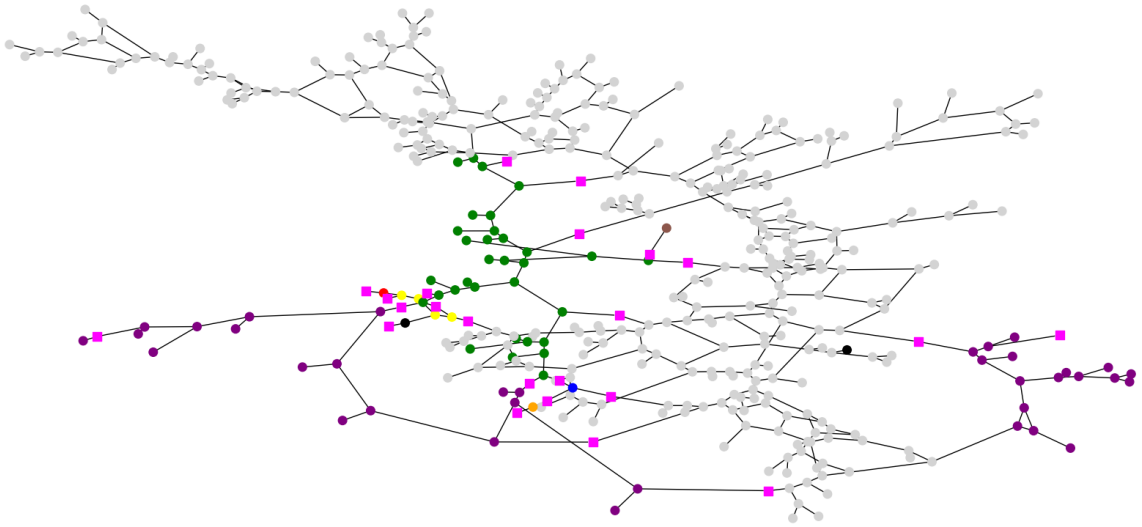


Figure A.17.: Pressure stages of instance Res2.

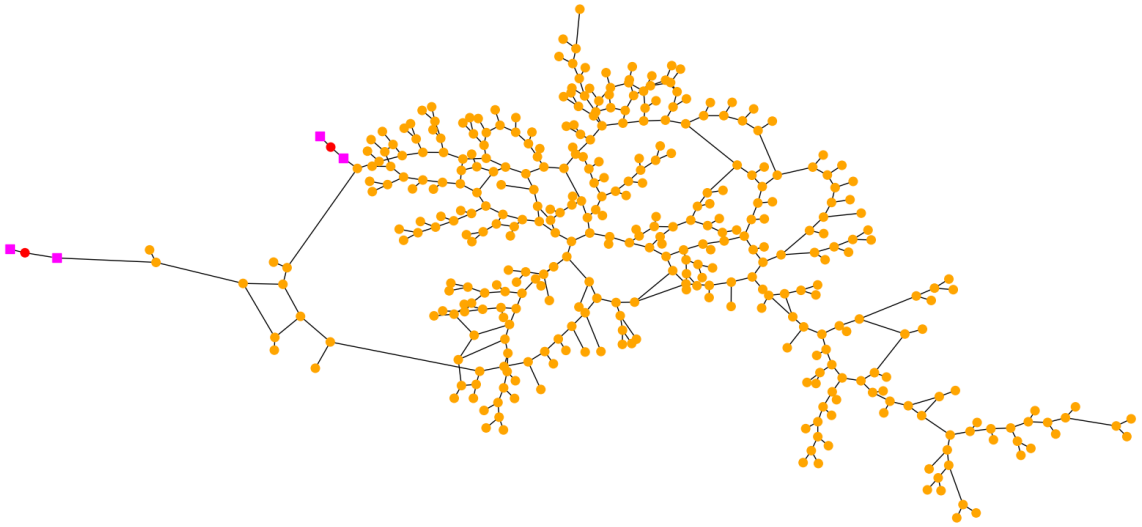


Figure A.18.: Pressure stages of instance Res3.

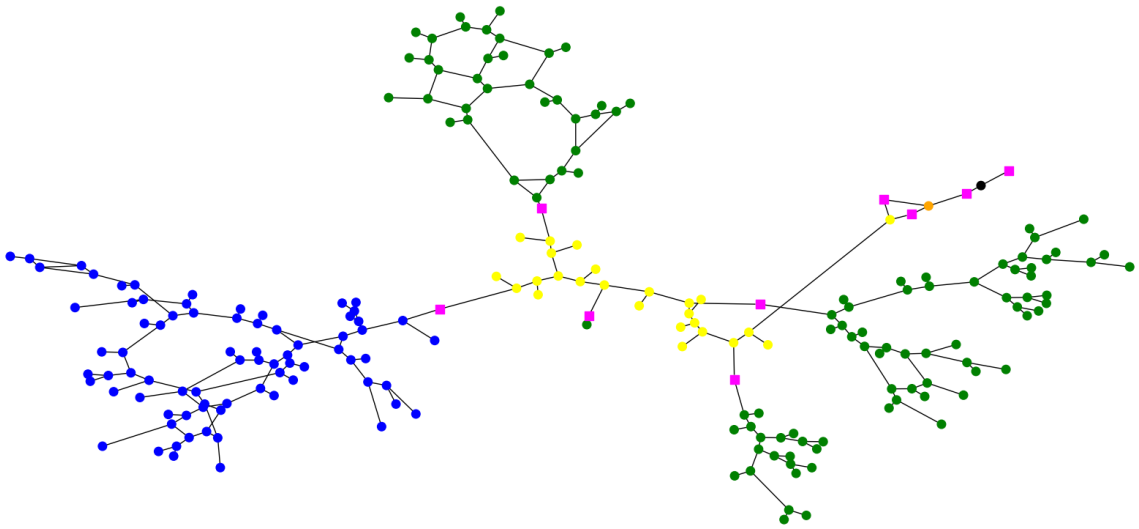


Figure A.19.: Pressure stages of instance Rur1.

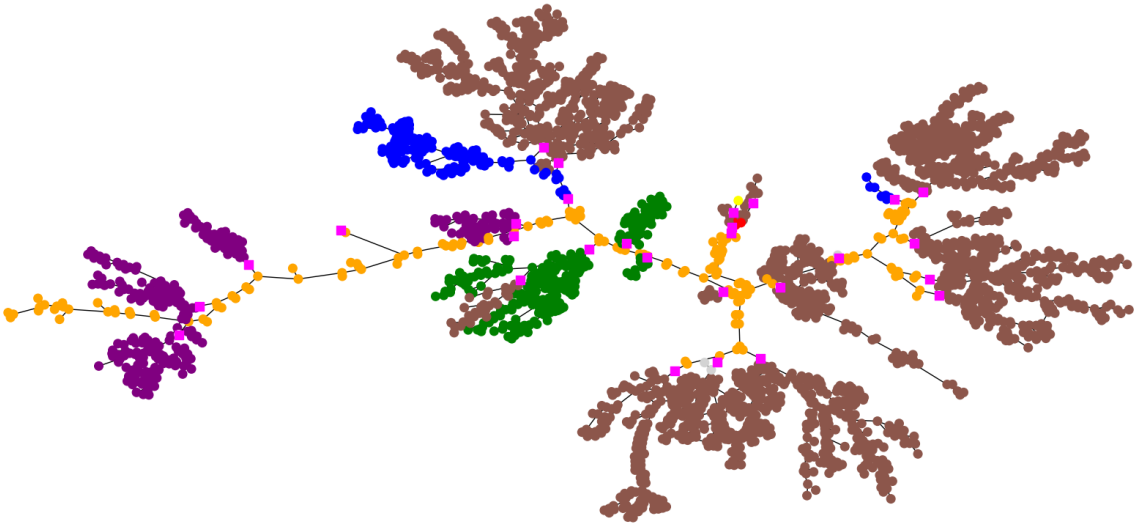


Figure A.20.: Pressure stages of instance Rur2.

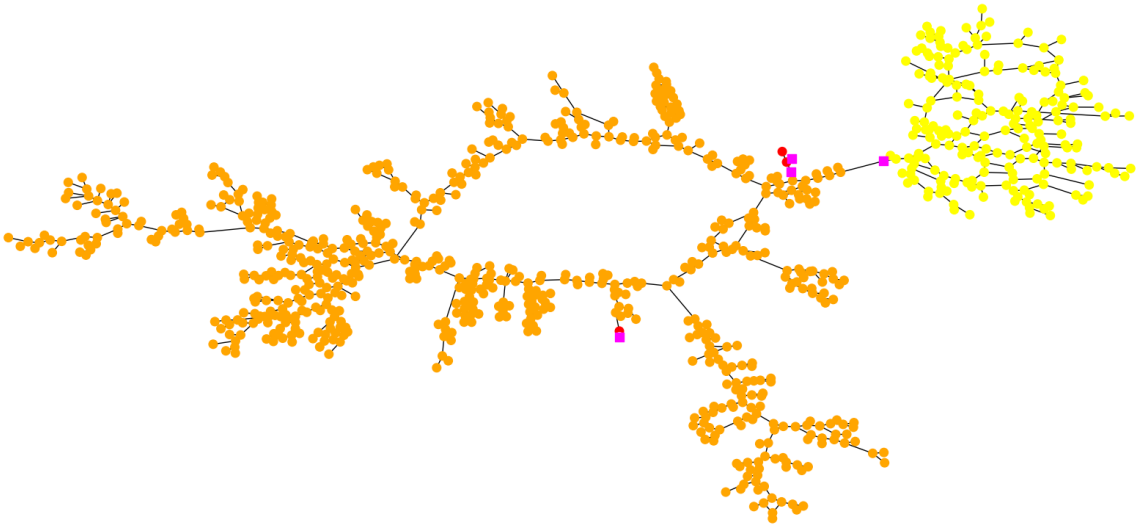


Figure A.21.: Pressure stages of instance Rur3.

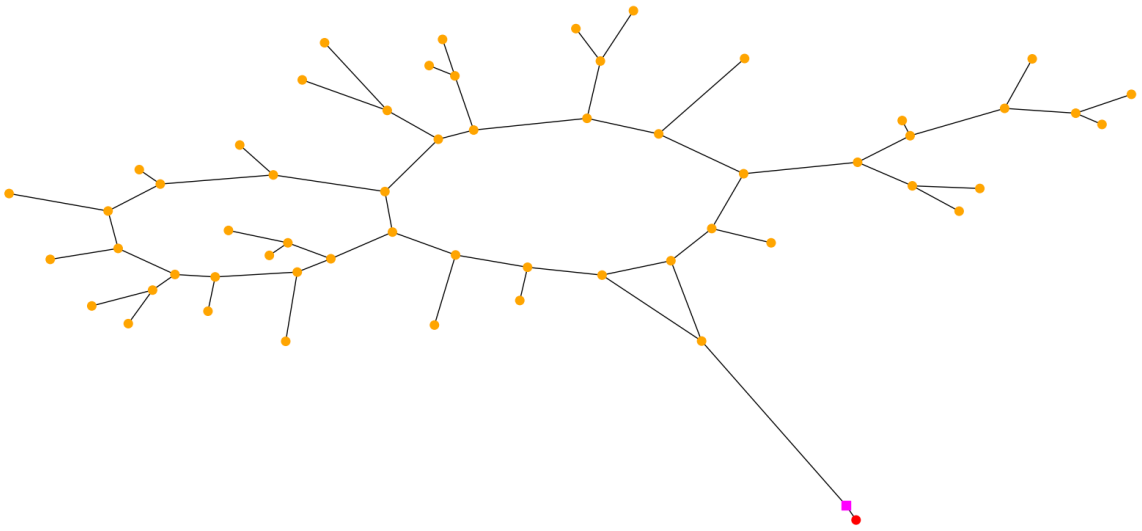


Figure A.22.: Pressure stages of instance Ind1.

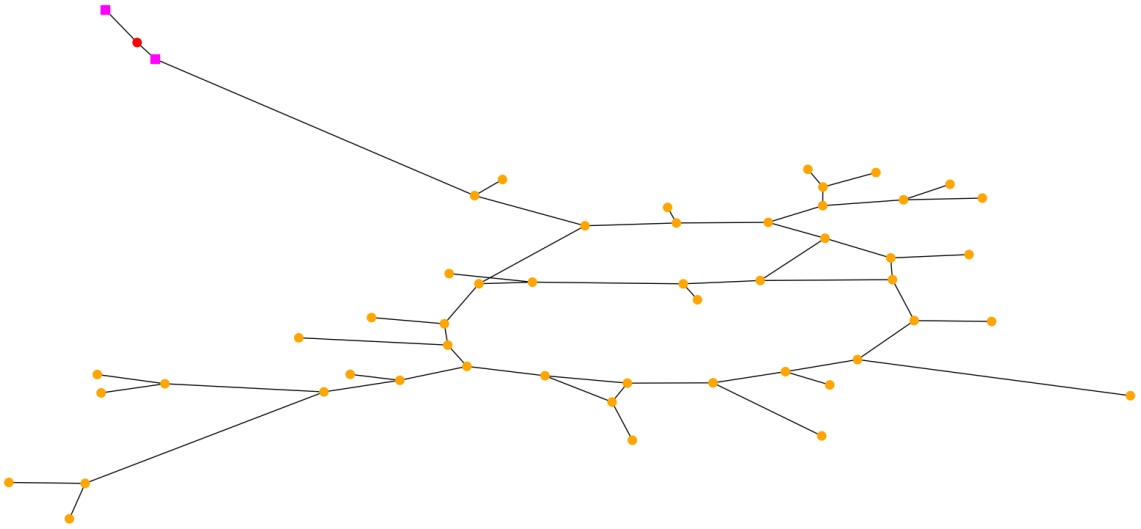


Figure A.23.: Pressure stages of instance Ind2.

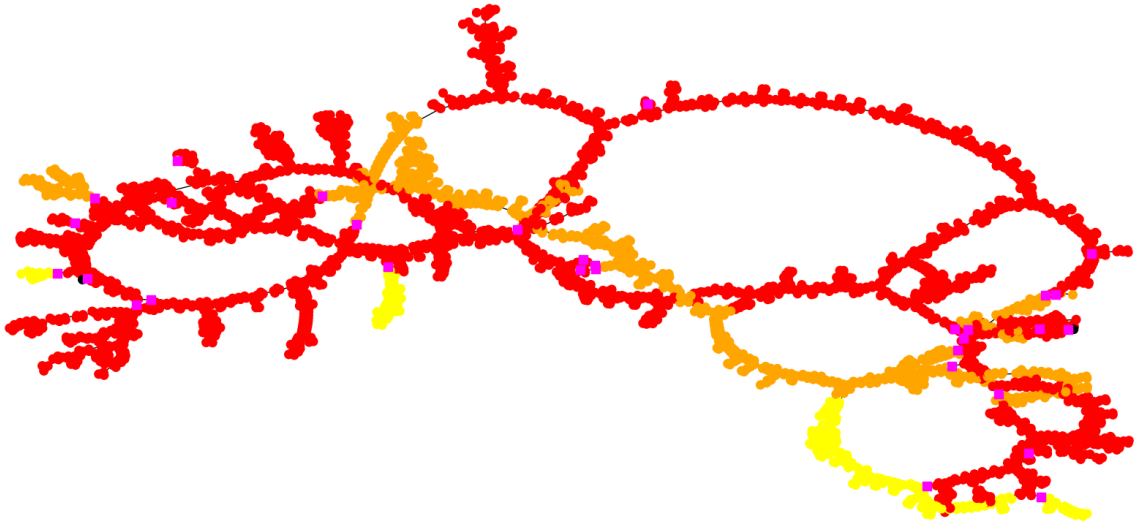


Figure A.24.: Pressure stages of instance Sup1.