Building a Data Pipeline to Analyse symmetries in Network

Jai Prathap Gomathi Veerakumar
and 201570386

Supervised by Jonathan A. Ward

Submitted in accordance with the requirements for the
module MATH5872M: Dissertation in Data Science and Analitics
as part of the degree of

Master of Science in Data Science and Analystics

The University of Leeds, School of Mathematics

September 2022

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Abstract

A network or graph is a collection of vertex points and a set of edges connecting those vertex points in pairs. Numerous phenomena, such as how people influence one another's voting decisions and the transmission of illness during an epidemic, can be modelled as dynamical processes on networks. The large networks are analysed by the method of reducing the network symmetries. Network symmetries, which emerge from the combinations of vertices without affecting the network's structure, can be used to reduce the amount of computing required to study such dynamics on networks with accuracy [18]. Graph theory, Group theory and Markov Chain are the mathematical concepts behind the network symmetry analysis. In this project, I am using a data set that comprises all graphs up to 9 vertices, which includes more than a million networks, to study network symmetries. The project is based on building a pipeline to analyse these data by reducing the symmetrical analysis. I will discuss my efforts to handle and analyse these data in this report.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Influence is a method of affecting someone's behaviour or decision. Politics and campaigning is a game of chess with influence playing a crucial role (figure 1.1). This chess is not about destroying the opponent's supporters, but it is about attracting them. The decision of the majority of supporters in this game determines the winner, but most of the supporters are under the influence of someone else. One of the complex tasks of the players in this game is the calculation of their influence on the majority.

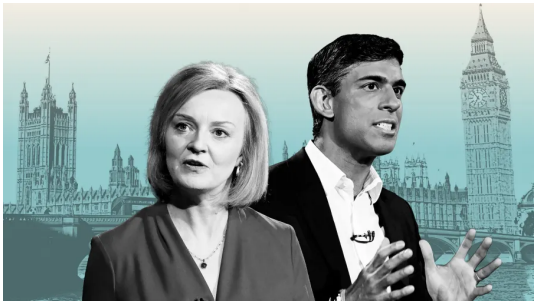

Figure 1.1: Uk prime minister campaign [13]

Like politics calculation of the spread of disease in an epidemic is a complex analytical task. Covid-19 has proven and shown the world about complex analysis (figure 1.2). These complex

Figure 1.2: Covid epidemic [4]

analyses are considered a dynamic process on a network. Dynamic processes are the influence of a vertex on another vertex or a group of vertices of the network [18]. The small-scale networks are analysed exactly, as the number of permutations and combinations for a small-scale network is less. The larger-scale networks are analysed by exploiting the method of redundancies [18]. The symmetries in the permutations method are reduced to have an approximation analysis of the network [18]. Symmetries are different combinations of a network without affecting the network's structure [18]. Graph and Group Theory are the mathematical concepts behind the analysis of the symmetries in a network. The GAP (Group Algorithm Programming) software analyses different properties of the network. The GAP analyses the mathematical computations of data. Python and its Libraries can analyse the statistical properties of the network. The computation requires fats process and high storage memory, and these are offered by HPC (High-Performance Computing). There are various steps to analyse the symmetries in the network, and these methods are combined to form a pipeline. The various steps in the pipeline are Data Extraction, Installation of software in HPC, Conversion of graph files to CSV and Statistical and symmetry analysis. This report summarises the different types of networks and the pipeline built to analyse these symmetries in any network.

Chapter 2

Mathematical Background

2.1 Network or Graph

Networks or a graph (G) is a combination of 2 or more vertices with edges and links [12], and it is represented by ($G = (V, E)$) [19] . Vertex or node are the edges of a connection in a network [18]. Several combinations of networks are formed from factors like the number of vertices, inter-connectivity and structure. The connectivity between the vertices forms the structure of a network. Table 2.1 represents the number of different graphs that can be formed in given vertices.

Vertices	Number of Graphs
2	2
3	4
4	11
5	34
6	156
7	1044
8	12346
9	274668
10	12005168
11	1018997864

Table 2.1: Number of graphs in each vertices [10]

The interconnectivity between the vertices forms the structure which in terms defines the formation of symmetries. There are different types of interconnectivity between the vertices:

- Null Graph
- Regular Graph
- Complete Graph

- Cyclic Graph
- Simple Graph

2.1.1 Null Graph

A graph in which there is no interconnectivity between the vertices is called a Null Graph [2]. The null graph can be rotated or mirrored without affecting the graph's structure.

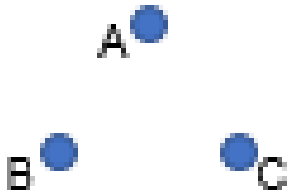

Figure 2.1: Null Graphs

Figure 2.1 represents a three vertex null graph.

2.1.2 Complete Graph

A graph, in which all the vertices are connected to each other is called a complete Graph [17]. The figure 2.2 shows the complete graphs for 2, 3, 4, 5 and 6 vertices. The interconnectivity is

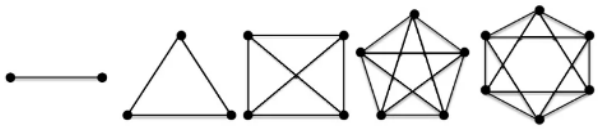

Figure 2.2: Complete Graphs [17]

formed between all the vertices in a graph [17].

2.1.3 Regular Graph

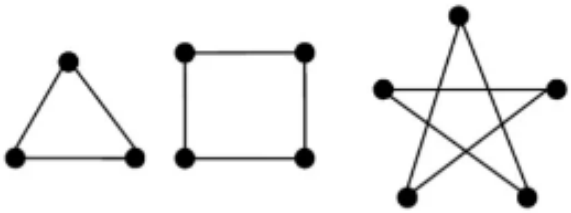

Figure 2.3: Regular Graphs [17]

A graph in which all the vertices are in equal degree is called a regular Graph [17].
The figure 2.3 represents the regular graph for three, four and five vertices. The three vertices regular graph is a triangle with an equal degree of 60.

2.1.4 Cyclic Graph

A graph in which all the vertices have only two interconnectivity links is called a cyclic graph [17]. Apart from the null graph Cyclic group has the second highest automorphism group. The

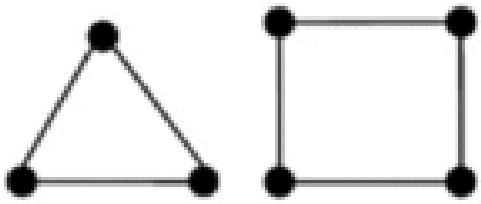

Figure 2.4: Cyclic Graphs [17]

figure 2.4 represents a cyclic graph. The state space for two infected people always moves in a pair.

2.1.5 Simple Graph

A simple graph is a combination of different graph categories. It can be with or without connectivity.

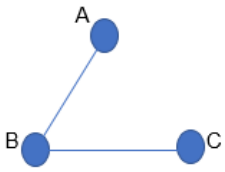

Figure 2.5: Simple Graphs

In this graph (figure 2.5), there is a connection between A B and B C, but there is no connection between AC.
It is hard to analyse symmetries in a simple graph due to the complexity of the connections between the vertices. The different states of the cyclic graph are easy to analyse as the interconnectivity is easy to understand.

2.2 Network Structure

The networks are represented in different structures, and the number of symmetries formed can vary from structure to structure. The most common structures in the network (figure 2.6):

- line
  The connections between the vertices make it a line. The automorphism groups are formed by reflection and identity (figure 2.6) [20].

- Ring
  The connections between the vertices make it a ring [20]. The automorphism groups are formed by reflection, rotation and identity (figure 2.6).

- Mesh
  The mesh structure is one of the complex structures. The automorphism groups can only be formed by identity due to their complex structure (figure 2.6).

- Star
  In a star graph, all the vertices are connected to only one vertex, which makes it look like a star (figure 2.6) [20]. The automorphism groups can be formed by reflection, rotation and identity.

- Tree
  The tree structure occurs when a vertex has other vertices as its branch and leaves (figure 2.6) [20]. The automorphism groups can form identity and sometimes by the reflective method.

- Bus
  The bus structure occurs when the connection between two vertices has branches (figure 2.6) [20]. The automorphism groups can form identity and sometimes by the reflective method.

- Fully Connected
  In a fully connected structure, all the vertices are connected to all the other vertices (figure 2.6) [20]. The automorphism groups are formed by reflection, rotation and identity.

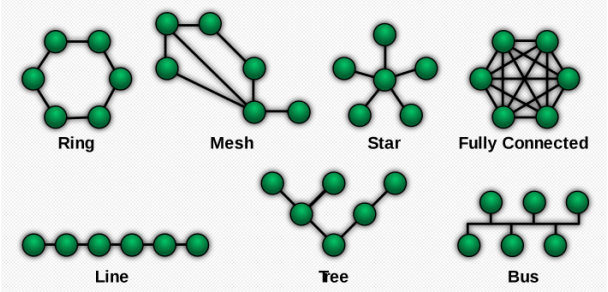

Figure 2.6: Different types of Network Structure [20]

The figure 2.6 represents all the different structures.

2.3 Symmetries and Automorphism

Symmetries or automorphism are permutations of vertices or nodes without affecting the network structure [18]. Null graphs and fully connected graphs have the highest number of symmetries. The mesh graph has the least number of symmetry.

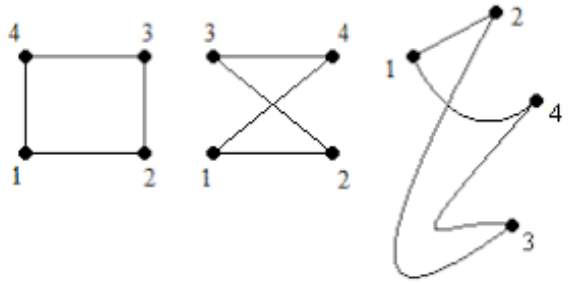

Figure 2.7: Different shapes of the nodes with same structure [7]

The structure consists of the connections formed between the nodes. In the figure 2.7, the graph's shapes are different, but the connectivity between the vertices remains the same, and its structure is unaffected.

2.3.1 Automorphism in a Cyclic Graph

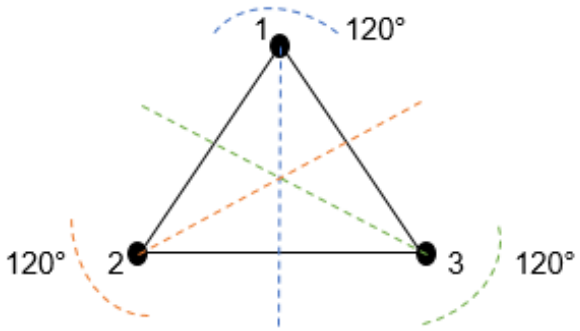

Figure 2.8: Three vertex complete graph

It is a three vertex complete graph (figure 2.8), which in normal words is a triangle with vertices 1, 2, 3. Vertex 1 is made stationary, and the graph is flipped such that the position of the vertex is now filled by vertex 3 and the position of vertex 3 is filled by vertex 2 (figure 2.8). The flipping of the graph is represented by the blue dotted line (figure 2.8). The flipping of the graph is also known as the reflection [9]. The position of the vertex changes, but the structure of the graph remains unchanged. It can represented as [9]

$$1 \longrightarrow 1$$

$$2 \longrightarrow 3$$

$$3 \longrightarrow 2$$

In mathematical terms, 1 is mapping itself, 2 is mapping 3, and 3 is mapping 2. This type of mapping is known as the permutation mapping, it maps the vertex set of a graph to the vertex set of the same graph [9]. It is represented by the equation

$$\alpha : V(G) \longrightarrow V(G)[9] \tag{2.1}$$

$$\alpha = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} = (1)\ (2\ 3) \tag{2.2}$$

From the equation 2.2, the notation (2 3) tells that the permutation only happens between 2 and 3 in this symmetry. If a single vertex is written in this notation (1) then that vertex is fixed. Another way of analysing automorphism is by rotating the graph 120° either clockwise or anti-clockwise [9]. Identity is also an automorphism group where all the vertices of graph G map themselves [20]. This identity automorphism is the only automorphism that is common in all the graphs [20]. The different automorphism groups of this graph[9]:

- Reflection 1, the vertex 1 is fixed [9]

$$1 \longrightarrow 1$$

$$2 \longrightarrow 3$$

$$3 \longrightarrow 2$$

$$\alpha 1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} = (1)\ (2\ 3) \tag{2.3}$$

- Reflection 2, the vertex 2 is fixed [9]

$$1 \longrightarrow 3$$

$$2 \longrightarrow 2$$

$$3 \longrightarrow 1$$

$$\alpha 2 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} = (2)\ (1\ 3) \tag{2.4}$$

- Reflection 3, the vertex 3 is fixed [9]

$$1 \longrightarrow 2$$

$$2 \longrightarrow 1$$

$$3 \longrightarrow 3$$

$$\alpha 3 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} = (3)\ (1\ 2) \tag{2.5}$$

- Rotation 1 (Clock wise) [9]

$$1 \longrightarrow 2$$

$$2 \longrightarrow 3$$

$$3 \longrightarrow 1$$

$$\beta 1 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} = (1\ 2\ 3) \tag{2.6}$$

- Rotation 2 (Anti Clock wise) [9]

$$1 \longrightarrow 3$$

$$2 \longrightarrow 1$$

$$3 \longrightarrow 2$$

$$\beta 2 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} = (1\ 2\ 3) \tag{2.7}$$

- Identity [9]

$$1 \longrightarrow 1$$

$$2 \longrightarrow 2$$

$$3 \longrightarrow 3$$

$$I = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} = (1)\ (2)\ (3) \tag{2.8}$$

2.3.2 Automorphism in Asymmetrical Graph

The graphs without symmetries are called Asymmetrical Graphs. Most of these graph belongs to the class of line or mesh graph.

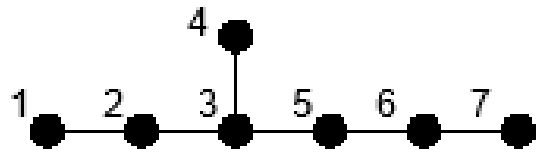

Figure 2.9: Seven vertex asymmetrical network [14]

The graph in figure 2.9 is an example of an Asymmetrical graph. The only possible automorphism group is identity.

$$1 \longrightarrow 1$$

$$2 \longrightarrow 2$$

$$3 \longrightarrow 3$$

$$4 \longrightarrow 4$$

$$5 \longrightarrow 5$$

$$6 \longrightarrow 6$$

$$7 \longrightarrow 7$$

$$I = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{pmatrix} = (1)\ (2)\ (3)\ (4)\ (5)\ (6)\ (7)$$

The graph (figure 2.9) has seven vertices, and no symmetries could be obtained due to the connection between vertex 3 and 4.

The following conclusions can be obtained for the automorphism group:

- If α and γ are the automorphisms of a graph (G), then $\alpha^{\circ}\gamma$ is also an automorphism of the graph [9].
- An identity map is an automorphism group for all graphs [9].
- If α is an automorphism of graph G, then the inverse (α^{-1}) is also an automorphism of graph G [9].
- $\therefore$ The set of G forms a group under the composition operation [9].

2.4 State Space Orbits

The orbits are the set of vertices that are identical to the given vertex. The orbits analyses the number of identical vertex in each orbit, and if the number of orbits is equal to the number of vertices, then there is no chance of building symmetry [18]. The vertices in a graph can be represented by the notation V [18]. The five vertices in the graph are represented by

$$V = \{1, 2, 3, 4, 5\} \tag{2.9}$$

Let's assume that S is susceptible and I am infected. The two possible states that a vertex can have is S or I. The state of the vertex is known as Vertex State [18]. The vertex state can be represented by the symbol W [18].

$$W = \{S, I\}[18] \tag{2.10}$$

The state space of the vertices can be represented by the symbol Ω [18].
The state space can be represented as

$$\Omega = W^V [18] \tag{2.11}$$

Let us consider a function f in the state space Ω [18].

$$f(u) \in \Omega [18] \tag{2.12}$$

Let u be the vertex in the set vertices V [18].

$$u \in V [18] \tag{2.13}$$

From the equation 2.10, if the state space of the five vertex is $\{\text{SSSSI}\}$. The state of the individual vertex can be represented as

$$f(1) = S, \;\; f(2) = S, \;\; f(3) = S, \;\; f(4) = S, \;\; f(5) = I,$$

All the different possible states of the vertex combined together forms a state space orbit.

2.4.1 Example of state Space Orbit

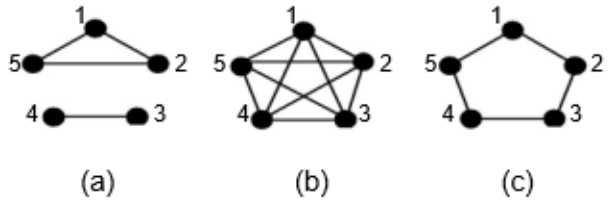

Figure 2.10: Different types of five vertex network [3]

In the figure 2.10, the number of orbits that can be formed is 2. The vertices 1, 2, and 5 form the first orbit. Vertices 3 and 4 form the second orbit. The 1, 2 and 5 can swap among themselves and they are independent of the vertex 3 and 4. Vertex 3 and 4 can swap among themselves and they are independent of the other three vertices. The vertex orbits of the graph in the figure 2.10a are represented as

$$VertexOrbits = [\{1, 2, 5\}, \{3, 4\}] \tag{2.14}$$

From the equation 2.14, there is no connection between the first and second orbits, which makes both the orbits independent of each other.
From the equation 2.14, the state Space orbits for 1 infected individual can be represented as

$$\Omega = [\{ISSSS, SISSS, SSSSI\}, \{SSISS, SSSIS\}]$$

The permutation is applied to both the vertex orbits and state space orbits to find all the state spaces of the graph.
In the figures 2.10b and 2.10c, the number of orbits that can be formed is 1. The vertices 1, 2, 3, 4 and 5 form single separate orbits. The vertex orbits of graphs in figures 2.10b and 2.10c are represented as

$$V = \{1, 2, 3, 4, 5\}[18] \tag{2.15}$$

From the equation 2.15, the state Space orbits for 2 infected individuals in the network (figure 2.10b) can be represented as

$$\Omega = [\{IISSS, ISISS, ISSIS, ISSSI, SIISS, SISIS, SISSI, SSIIS, SSISI, SSSII\}$$

From the equation 2.15, the state Space orbits for 2 infected individuals in the network (figure 2.10c) can be represented as

$$\Omega = [\{IISSS, SIISS, SSIIS, SSSII, ISSSI\}$$

The graph (figure 2.10) can only rotate as it is a cyclic graph and all the vertices are fixed to two other vertices, so the vertex state for two infected moves in a pair.

2.5 Markov Chain

From the paper Dimension-reduction of dynamics on real-world networks with symmetry by Jonathan A. Ward, Markov chain describes the dynamical processes on finite networks in which only one vertex can change vertex state and each vertex can be in one of a finite number of vertex states [18]. The total number of state space vertices of N vertices is calculated by the formula

$$M = 2^N[18] \tag{2.16}$$

Let S be the state space which has multiple state $S = \{s_1, s_2, ..., s_{M^N}\}$ [18].Assume that the relevant time-dependent probability distribution is $X(t) = (x1(t), x2(t), ..., xMN(t))^T$ over a state s, then the $X(t)$ is depicted by forward Kolmogorov [18]

$$X(t) = Q^T X[18] \tag{2.17}$$

where Q is an $M^N \times M^N$ matrix called the infinitesimal generator whose ij^{th} component describes the transition rate from the state s_i to the state s_j for i $\neq$ j [18]. Markov Chain is the method to find the probability state of the future event from the state of the previous event [15]. The probability of obtaining the state of vertex 3 in figure 6 depends upon the state of vertex 2. Markov chain can be represented as:

$$P(future|past, present) = P(future|past, Markovproperty.)[15] \tag{2.18}$$

A finite Markov chain is lumpable if there is a partition of state-space $L = \{L_1, L_2, ..., L_r\}$ on which the Markov property is retained and it has been demonstrated that models can be lumped using network symmetries [18], [19]. Markov property is a value that makes the calculation much easier [15]. It states that the probability of the future state at a given time depends only on the present state, not on all past states [15]. By lumping states collectively and exploiting network symmetries to shrink the state-space, it is possible to do precise analysis of larger networks [18]. All the concepts of the Markov Chain section were fully based on the reference from the paper Dimension-reduction of dynamics on real-world networks with symmetry by Jonathan A. Ward.

2.6 Complement of the Graph

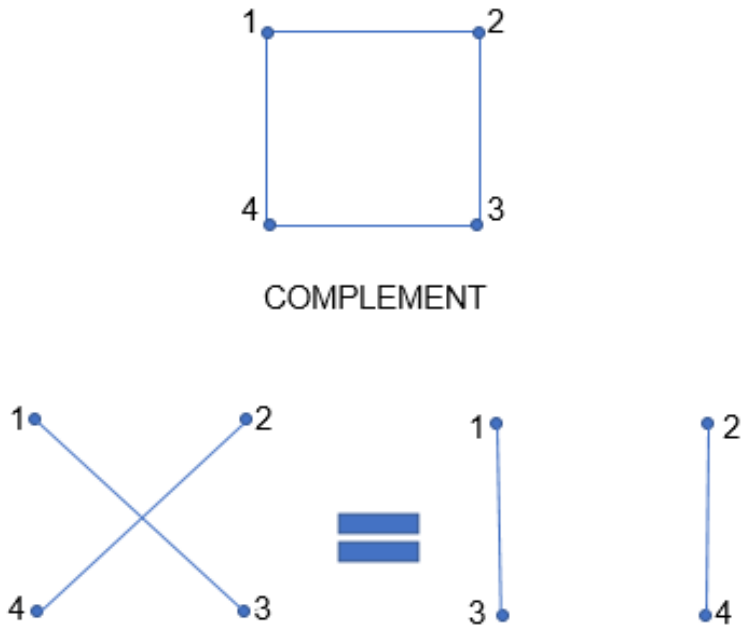

Figure 2.11: Four vertex graph and its complement

Graphs are with a particular number of vertices, and the edges represent the interconnectivity between the vertices [8]. The graph's complement is such that the connection between the edges only exists at places where there are no connections between the edges in the original graph [8]. Let us assume a graph G with vertices v and edges e (G(v, e)) [8]. The graph's complement with vertices v and the edges e' (G'(v, e')) [8]. In the figure 2.11, the connection between the edges of the square network are (1,2), (2,3), (3,4), (1,4). The connections between the edges of the complement graph are (1, 3) (2, 4) (figure 2.11). The connections (1,3) (2,4) does not exist in the square graph (figure 2.11). The diagonal connections can also be drawn in another way (figure 2.11), where the shape of the graph differs but not the structure. The complement of the graph helps to understand the complexity of the structure. More complex structure's compliment is

simple, so the complexity of the structure can be analysed.

2.7 Application of Automorphism

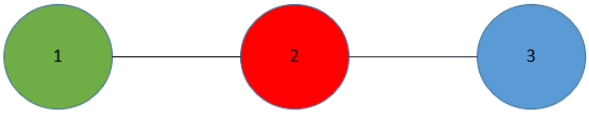

Figure 2.12: A simple three vertex network

Figure 2.12 represents a network between three persons: person1, person2 and person3. During an epidemic, there are two possibilities for a person to have either infected or susceptible (equation 2.10). There are eight possible outcomes that this network can have, and these eight possible outcomes combined are known as the state space of the network.
Let's assume that S is susceptible and I am infected. The two possible states that a vertex can have is S or I (equation 2.10). The possible outcomes are known as state space, which is clearly explained in the section state space orbits. All eight possible outcomes are represented as

$$\Omega = \{SSS, ISS, SIS, SSI, SII, ISI, IIS, III\} \tag{2.19}$$

Reflection , the vertex 2 is fixed

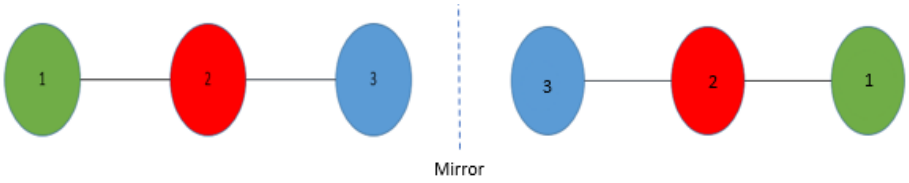

Figure 2.13: Reflection of three vertex network

$$1 \longrightarrow 3$$

$$2 \longrightarrow 2$$

$$3 \longrightarrow 1$$

$$\alpha = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} = (2)\ (1\ 3)$$

From the equation 2.19, due to the symmetries, ISS and SSI are considered the same. If the network is mirrored the ISS outcome looks like SSI (figure 2.13). There are two possible symmetries in the outcome:

$$(ISS = SSI, SII = ISS)$$

The analysis reduces possible outcomes from eight to six, which reduces computational errors and storage space. The symmetries play a crucial role in computing large graphical data by reducing the computational memory. Table 2.1 interprets that there is an exponential increase in the number of graphs to the vertices number. Calculating the symmetrical graphs is identical to calculating the same graph multiple times. The chance of having more symmetry is high if the number of automorphism groups is large.

Graph and Group Theory are the mathematical concepts used to analyse the auto morphism of graphs. An automorphism of the graphs is carried out to analyse the symmetries in the network to reduce the dimension of the network [15].

Chapter 3

Pipeline

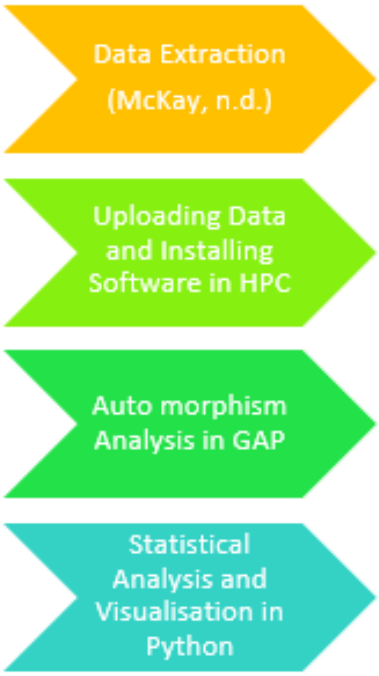

Figure 3.1: Pipeline for the analysis of network

The pipeline is a sequential combination of several processes, from Data Extraction to statistical analysis (figure 3.1). The process is a sequence and related, and it is also a part of Markov Chain analysis: the analytical input for the future state (process) depends on the output of the present state (process). Figure 3.1 represents the pipeline for analysing the data.

3.1 Data Extraction and Description

Data is extracted from Brendon McKay's website [10]. The website provides the data file for vertices 2 to 11, and the source data file is in the form of a graph file (.g6). The graph data

format is used for storing undirected graphs in a compressed manner [5]. These graphs have ASCII codes, which are unique for all the graphs (figure 3.2) [5].

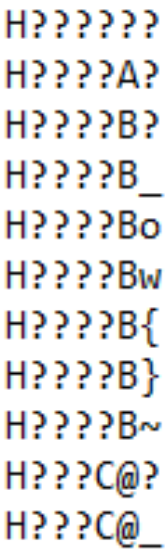

Figure 3.2: ASCII values of nine vertex G6 file [10]

Figure 3.2 represents the ASCII Code for vertex 9. The ASCII code contains the necessary description, including the image of the graph.

Vertices	ASCII Code Starting Letter
2	A
3	B
4	C
5	D
6	E
7	F
8	G
9	H
10	I
11	J

Table 3.1: First letter of Code for each vertex [10]

The first letter of each code for different vertices starts from the alphabet given on the table 3.1. The first letter indicates the number of vertices. After a few hundred vertices, the number of vertices can be identified by the first two letters. If observed closely, there is a pattern in the code. The last digit of the first code in figure 3.2 is a question mark, and the last before digit in the first code is also a question mark. The last digit of the second, third and tenth code is a question mark (figure 3.2). When the second last digit changes, the last digit of the code starts

with a question mark (figure 3.2). The intriguing fact about the code is they are all unique. The

Repeated Codes
A?
B?
Bw

Table 3.2: The Repeated codes in vertex 9 [10]

ASCII codes in the table 3.2 are the codes designed for the graphs with 2 and 3 vertices. The last two digits of the second, third and sixth code in the figure 3.2 and the ASCII codes in the table 3.2 are the same. The makers of the code may have generated a pattern, and they have started to encode the data in that pattern or the repeated codes in higher vertex tells us the resemblance of both the graphs.

3.2 HPC

HPC (High-Performance Computer) is used to process data or software of large files. The University of Leeds provides High-Performance Computing facilities for analysing large files. The HPC is a Linux-based system with large memory and high processing speed. The python IDE anaconda is pre-installed in the HPC, but the GAP requires installation. The HPC can be operated from off campus, which is by the use of MobaXtem personnel. It provides interactive remote access from pc to the HPC. All the installations and processes in the software are by providing Linux command. The user interface is not friendly for a windows user.

3.3 Autmorphism Analysis in GAP

GAP (Group Algorithm and Programming) is a computation software used for discrete maths analysis focused on computational group theory [5]. GAP provides many functions, mathematical algorithms and libraries to analyse algebraic functions [5]. The graph file is analysed by utilising the package digraph. The digraph is a package developed by the University of St. Andrews to analyse mutable and immutable digraphs [5]. The interface and the code created are similar to C and C++ programming. Understanding the user interface and the commands is a bit complex due to the mix-up of small and capital letters, as most programming languages use only small letters.
The following properties of the graphs are extracted using the GAP software (figure 3.3):

- Structure Name
- Number of Vertices
- Order of Graphs
- State Space of the Orbits

- Description

- Vertex Orbits

The `DigraphNrVertices` command returns the number of vertices in a graph [5]. The input provided is a 9 vertices graph format file and this is done to avoid any misleading information in the input file. The `AutomorphismGroup` command returns all the automorphism groups of the graph [5]. It will be huge for the data file to process, so the order of the automorphism group is taken. The order of the automorphism group is the sum of the cardinality of the vertex of all the automorphism groups [5]. Cardinality is the number of the vertex in each group, it is represented as $|V|$ [18]. The state space of the orbits is obtained by using a function called `PolyaEnumeration` [18]. It calculates the state space of all the automorphism groups [18]. From the equation 2.16, for the vertices N= 9:

$$M = 2^9 = 512$$

If M is equal to 512 then the state space doesn't have any symmetries in them. The total state space orbits will be less than 512 if the vertex has symmetry in them. The vertex Orbits can be obtained from the `OrbitLengthsDomain` function [18]. The orbit length is to obtain the length of different orbits formed to understand the number of different orbits to get an idea about the structure. If we take the figure 2.10 as an example:

$$Orbits = [\{1, 2, 3\}, \{4, 5\}] = [3, 2]$$

The first orbit consists of three vertices and the second orbit consists of 2 vertices. The command `StructureDescription` returns the description of the graph [5]. For example: if a graph's description is S7, it depicts that the graph belongs to the star group with 7 vertex.

```
gap> fout:=OutputTextFile("E:/gap/graph9_data.csv",false);
OutputTextFile(E:/gap/graph9_data.csv)
gap> for graph in graphs do
> N:=DigraphNrVertices(graph);
> gaut:=AutomorphismGroup(graph);
> nORs:=PolyaEnumeration(gaut,N,2);
> s:=StructureDescription(gaut);
> gs:=Graph6String(graph);
> AppendTo(fout,";",gs,";",N,";",Order(gaut),";",nORs,";",s,";",OrbitLengthsDomain(gaut,DigraphVertices(graph)),"\n");
> od;
```

Figure 3.3: A clip of GAP Code

The input is processed in for loop, and all the required features are appended to a list separated by a semicolon delimiter (figure 3.3). The list is converted into a CSV file for statistical analysis in python (figure 3.3).

3.4 Statistical Analysis in Python

Python is the best tool for the data analysis process due to the availability of versatile packages to analyse the CSV file that was created in the GAP file.

3.4.1 Data Importing

```
df=pd.read_csv("gg9.csv",header=None,delimiter=';',
               names=["Index","Structure_Name","Vertices","Order","State_Space_Orbits","Description","Vertex_Orbits"],
               dtype={"Vertices":np.int64,"Order":np.int64,"State_Space_Orbits":np.int64},
               converters={"Vertex_Orbits": lambda x: list(map(np.int64,x.strip('[]').split(',')))})
```

Figure 3.4: Code for importing and modifying the CSV file

A normal csv file is comma separated file, but the delimiter of the csv file created by the GAP is semicolon. Semicolon delimiter is provided to the csv file is to avoid the library to identify the commas in the orbits as a delimiter. The header for the columns were not created in the GAP. The `read_csv` command enables the pandas to read csv and returns the dataframe (figure 3.4) [11]. The datatype of the vertices, order and state space orbits are converted into integer 64 (figure 3.4).
The map function transforms all the variables without iterating or using a for loop [11]. The vertex orbit length and the maximum number of vertices in a single orbit of the state space are created from the attribute vertices(figure 3.5). The number of asymmetrical graphs is identified

```
df['Vertex_Orbits_Length']=df.Vertex_Orbits.map(len)
df['Vertex_Orbits_Max']=df.Vertex_Orbits.map(max)
```

Figure 3.5: Code for creating vertex orbit length and max columns

by analysing the length of the vertex orbit. The only automorphism group of the asymmetrical graph is identity. In an asymmetrical graph, the list's length is equal to the total number of vertices.

$$Orbits = [\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{9\}] = [1,1,1,1,1,1,1,1,1] = 9$$

The code in figure 3.6 to determine the symmetry returns Boolean data. If the condition length vertex orbits are equal to 9, then the expression returns true. If the condition is not satisfied, then the value returns false.

Figure 3.7 represents the data frame used for statistical analysis.

```
df['Assymmetry']=df['Vertex_Orbits_Length']==9
```

Figure 3.6: Code for creating a Boolean column asymmetry

	Structure_Name	Vertices	Order	State_Space_Orbits	Description	Vertex_Orbits	Vertex_Orbits_Length	Vertex_Orbits_Max	Assymmetry
0	H??????	9	362880	10	S9	[9]	1	9	False
1	H????A?	9	10080	24	C2 x S7	[2, 7]	2	7	False
2	H????B?	9	1440	42	C2 x S6	[2, 6, 1]	3	6	False
3	H????B_	9	720	48	S5 x S3	[3, 5, 1]	3	5	False
4	H????Bo	9	576	50	S4 x S4	[4, 4, 1]	3	4	False
...	...	...	...	...	...	...	...	...	...
274663	H]~v~z~	9	384	30	((((C2 x C2 x C2 x C2) : C2) : C2) : C3) : C2	[8, 1]	2	8	False
274664	H]~v~~~	9	288	40	C2 x S3 x S4	[6, 3]	2	6	False
274665	H]~~~~~	9	960	36	S5 x D8	[4, 5]	2	5	False
274666	H^~~~~	9	10080	24	C2 x S7	[2, 7]	2	7	False
274667	H~~~~~	9	362880	10	S9	[9]	1	9	False

274668 rows × 9 columns

Figure 3.7: Dataframe created from the CSV file

3.4.2 Data Cleaning

Data cleaning is the primary step in a data analysis process. There are possibilities of the dataset having impurities these impurities affect the process; and cause misleadingness in the analysis.

```
df.isna().sum()
```

```
Structure_Name          0
Vertices                0
Order                   0
State_Space_Orbits      0
Description             0
Vertex_Orbits           0
Vertex_Orbits_Length    0
Vertex_Orbits_Max       0
Assymmetry              0
dtype: int64
```

Figure 3.8: Total number missing values in each column with the code

The misleadingness caused by the missing values is huge. The statistical attributes of the data like mean, median and mode will have a huge deviation from the data. The figure 3.8 interprets that there are no missing values in the data. The outliers are ignored to the loss of information on networks.

3.4.3 Dataset analysis

The total memory taken by the dataset is 17.0MB (figure 3.9). There are three categorical, four numerical and 1 Boolean attribute (figure 3.9). The vertex orbits, which are in the form of a list are considered to be unsupported, but they can also be taken as categorical (figure 3.9). There are no duplicate rows and missing values.

Dataset statistics

Number of variables	9
Number of observations	274668
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	17.0 MiB
Average record size in memory	65.0 B

Variable types

Categorical	3
Numeric	4
Unsupported	1
Boolean	1

Figure 3.9: Statistical output of the dataset

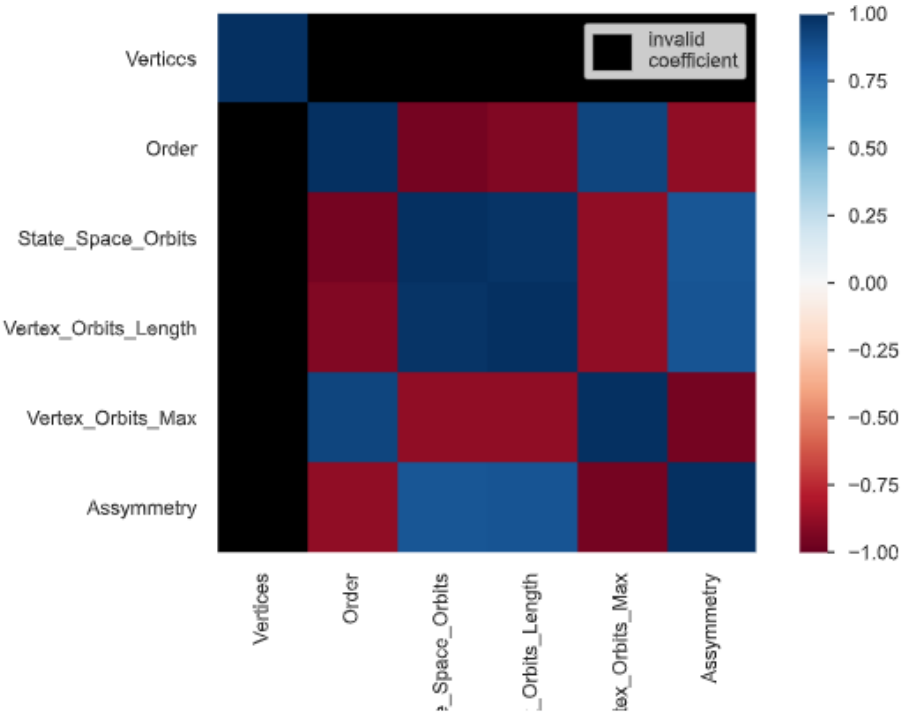

Figure 3.10: Correlation between all the variables

The correlation between the attributes is studied to understand the influence of all the attributes on each other(figure 3.10). The red colour represents the negative correlation between the two

attributes, which means the attributes are inversely proportional to each other (figure 3.10). The increase in the value of attribute A decreases attribute B. The blue colour represents the positive correlation between the two attributes, which means the attributes are inversely proportional to each other (figure 3.10). The increase in the value of attribute A with increase in the value of attribute B. The vertices attribute is unsupported according to the profiling report.
The following points are concluded from analysing the dataset attributes:

- Vertices have constant value of nine.
- Structure_Name has a high cardinality: 274668 distinct values due to the unique codes of the graph.
- The description has a high cardinality: 67 distinct values.
- Order is highly correlated with State Space Orbits (negative correlation) and three other fields(Asymmetry (negative correlation), Vertex Orbit Length (negative correlation), vertex orbit max).
- State Space Orbits are highly correlated with Order (negative correlation) and three other fields (Asymmetry, Vertex Orbit Length, vertex orbit max (negative correlation)).
- Vertex Orbits Length is highly correlated with Order (negative correlation) and three other fields (Asymmetry, State Space Orbits, vertex orbit max (negative correlation)).
- Vertex Orbits Max is highly correlated with Order and three other fields (Asymmetry(negative correlation), State Space Orbits (negative correlation), Vertex Orbit Length (negative correlation)).
- Asymmetry is highly correlated with Order (negative correlation) and three other fields (vertex_Orbit_length, State Space Orbits, vertex orbit max (negative correlation)).

Attribute Description Analysis

The first analysis is to understand the description details of the graph. The command `len` provides the sum of total number of variables in the list (figure 3.11). The graphs are sorted based

```
%matplotlib inline
bx=df['Description'].value_counts().head(5).plot(kind='bar',color='green')
bx.set_xlabel("Description")
bx.set_ylabel("Value Counts")
```

Figure 3.11: Code for plotting top 5 value counts of Description column

on the value count, and the command `head(5)` returns the top 5 values of the chart (figure 3.11) [11]. The analysis is depicted in a bar graph. Graphical visualisation is the best way to depict the data.

```
print(df.Description.unique())
len(df.Description.unique())
```

```
['S9' 'C2 x S7' 'C2 x S6' 'S5 x S3' 'S4 x S4' 'S3 x S5' 'S7' 'S8'
 'S5 x D8' 'C2 x C2 x S4' 'C2 x S3 x S3' 'C2 x S5' 'S3 x S6' 'C2 x S4'
 'S3 x S3' 'S5' 'S6' 'D8 x S3' 'C2 x C2 x S3' 'D12' 'S4' 'D8 x S5'
 'C2 x C2 x S5' '(S3 x S3) : C2' 'S4 x S3' 'C2 x D8' 'C2 x C2 x C2'
 'C2 x S4 x S3' 'C2 x S3 x S4' 'S3 x S4' 'C2 x C2 x D8' 'S4 x D8'
 'C2 x C2 x C2 x C2' 'C2 x C2' 'C2 x D8 x S3' 'C2 x C2 x C2 x S3' 'S3'
 'C2' 'S4 x S5' 'S4 x D10' 'D8' '1' 'D8 x S4' '((S3 x S3) : C2) x S3'
 'S3 x S3 x S3' '((((C2 x C2 x C2 x C2) : C2) : C2) : C3) : C2'
 'C2 x S3 x D8' 'D8 x D8' 'C2 x C2 x D10' 'C2 x ((S3 x S3) : C2)'
 'S5 x S4' 'D20' 'S3 x D10' 'D28' '(C2 x C2 x C2 x C2) : C2' 'D16'
 '(((C2 x C2 x C2 x C2) : C2) : C2) : C2' '(S4 x S4) : C2' 'D8 x D10'
 'D10' 'S6 x S3' 'D14' 'D18' 'D10 x S4'
 '((((C3 x C3 x C3) : (C2 x C2)) : C3) : C2) : C2' 'C3'
 'S3 x ((S3 x S3) : C2)']

67
```

Figure 3.12: Code and the unique values of Description column

There are 67 different groups of 9 vertices graph (figure 3.12). The command `unique` returns the unique values in the data frame [11]. There are three types of graph description with a different number of vertices:

- Star

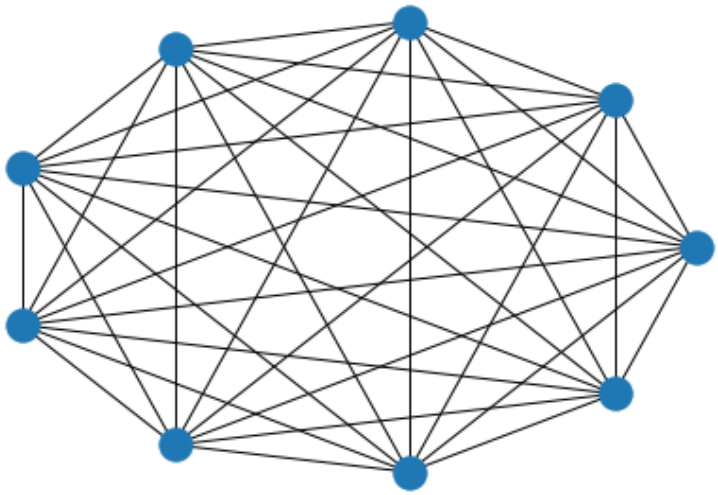

Figure 3.13: Star graph

The S9 is the star group with nine vertices (figure 3.13). These nine vertices are in the same graph. The second highest count of the graphs falls under the star group (figure 3.13). There are only two values which have the description name S9.

- Cyclic

The C2×C2is the two cyclic groups with two vertices each (figure 3.14). It is the highest

category in terms of the value count.

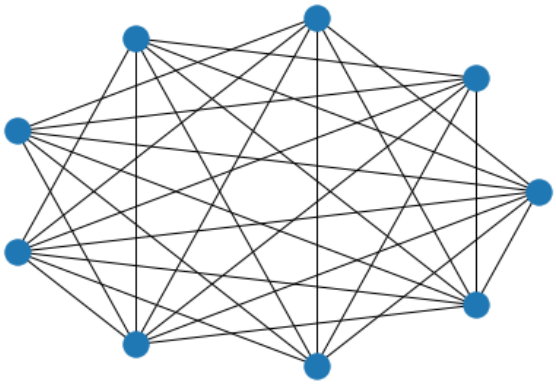

Figure 3.14: Cyclic graph

- Dihedral
  Dihedral is one of the complicated graphs (figure 3.15). n is the number of interconnectivity in the graph [18].

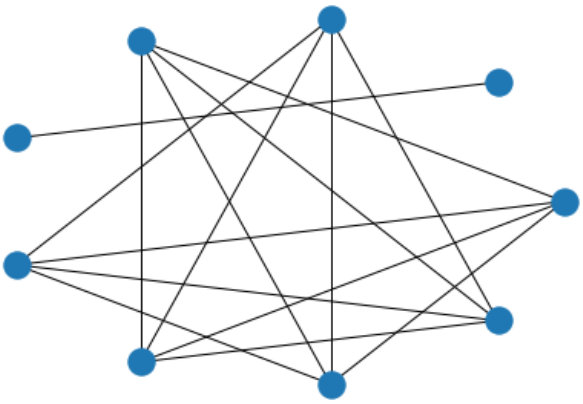

Figure 3.15: Dihedral graph

The formula of the description can be returned as X_n where x is the name of the group and n is the number of vertices in the group [18].
There are some groups like $((((C3 \times C3 \times C3) : (C2 \times C2)) : C3) : C2) : C2$ where the sum of the n values is greater than 9. The n represents the total number of vertices in that particular group. There are seven cyclic groups in the graph with some vertices repeating in different groups. There are some groups like S7 and S8 in 9 vertices graphs. S7 means there are seven vertices in the star group with two vertices in the null group. There are some notations like D28, D18, D14, D20 and D12. The n in the star and the cyclic group is the number of vertices, but the n in the dihedral represents the number of interconnectivity in the graphs [7]. The congruence

of the groups D20 and D28 can be written as:

$$D20 \cong C2 \times D10$$

$$D28 \cong C2 \times D14$$

The graph interprets the top 5 graph categories with a high-value count (figure 3.16):

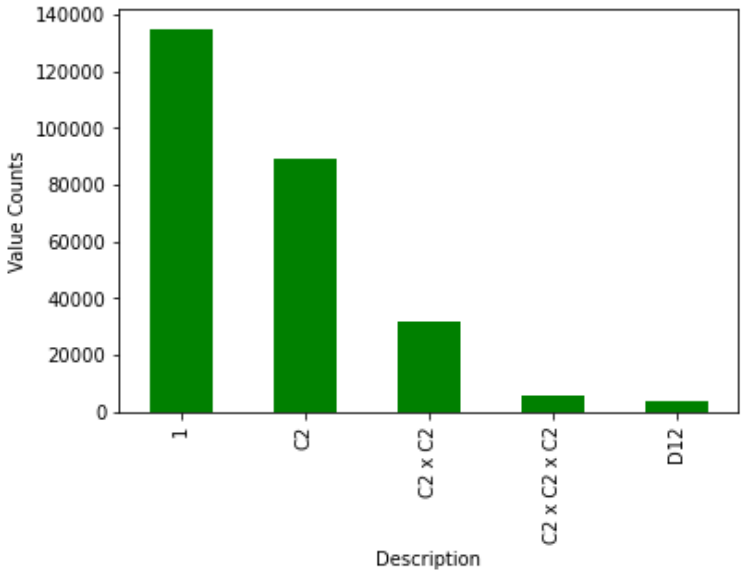

Figure 3.16: Bar graph of top 5 values

- 1
- C2
- $C2 \times C2$
- $C2 \times C2 \times C2$
- D12

Category 1 has the highest number of a graph in them. Around 50% of the graph falls into the category of 1 (figure 3.16). More than 75% of the graphs falls in the category of 1 and C2. The rest of the other groups from the remaining 20%.

```
bx=df['Description'].value_counts().tail(10).plot(kind='bar', color='magenta')
bx.set_xlabel("Description")
bx.set_ylabel("Value Counts")
```

Figure 3.17: Code for producing last 5 values

The command `tail(10)` returns at least ten value counts in the description (figure 3.17) [11]. The plot function draws a graph the attribute “kind” and “colour” depicts the type and colour of the chart (figure 3.18) [11].

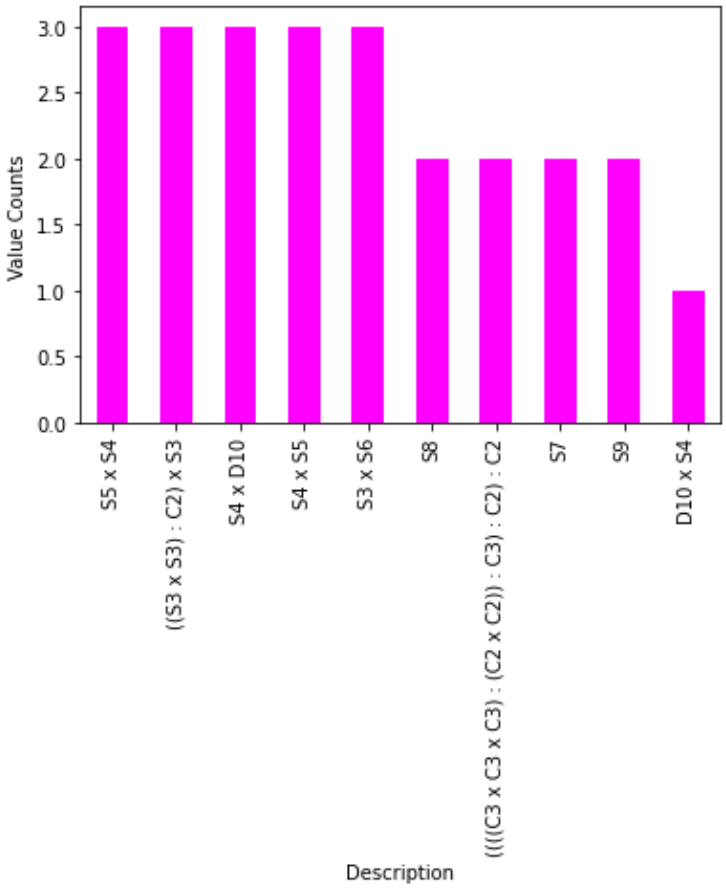

Figure 3.18: Bar graph of last 5 values

There is only one graph in the category D10xS4. The groups S9, S8, S7 and ((((C3 x C3 x C3): (C2 x C2)): C3): C2): C2 have only two graphs (figure 3.18). There are many groups with single-digit graphs in them. The analysis of the groups was done to estimate the complexity in the analysis of the graphs in the vertices. The complex graphs such as ((((C3 x C3 x C3): (C2 x C2)): C3): C2): C2 can take more memory in the analysis compared to the simpler graphs (figure 3.18). The groups with least five value counts (figure 3.18):

- D10 $\times$ S4
- S9
- S7
- S8

- ((((C3 × C3 × C3): (C2 × C2)): C3): C2): C2

Most occurring characters

Value	Count	Frequency (%)
2	181017	26.9%
C	177059	26.3%
1	138992	20.7%
	101688	15.1%
x	50709	7.5%
S	6938	1.0%
D	6511	1.0%
3	5796	0.9%
8	2487	0.4%
4	982	0.1%
Other values (8)	745	0.1%

Figure 3.19: Most occurring characters in the Description column

Character C has the highest among the three groups (figure 3.19). The 117069 networks are under cyclic group (figure 3.19). 6938 graphs are under star group and 6511 graphs falls under dihedral (figure 3.19).

Attribute Order Analysis

The command `value_counts()` returns the value count of any assigned column in the data frame [11]. The column or the attribute can be either numerical or categorical, but this command treats the numerical data as categorical data. The order is calculated to analyse the number of asymmetries and symmetries of different groups.
The order with top 5 value counts (figure 3.20):

- 1
- 2
- 4
- 8
- 12

Figure 3.20 interprets that the graphs with order 1 mean that around 50% of the graphs are asymmetric. Order is the sum of the cardinality of all the automorphism groups, and if the order is equal to one, then the graph is asymmetric [18]. The graphs with symmetries are separated

```
cx=df['Order'].value_counts().head(10).plot(kind='bar', color='blue')
cx.set_xlabel("Order")
cx.set_ylabel("Value Counts")
```

```
Text(0, 0.5, 'Value Counts')
```

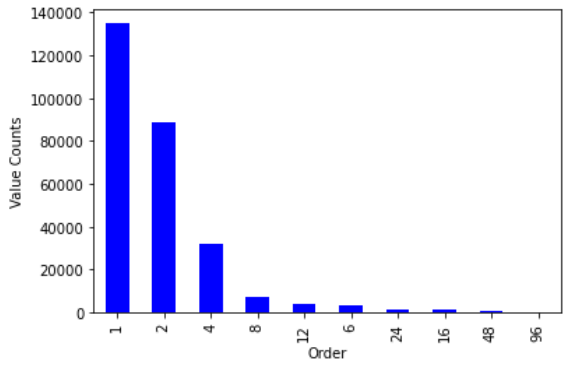

Figure 3.20: Code and bar graph of top 10 value counts in the column order

based on their order value. Around 75% of the graphs falls into the category of either 1 or 2. The more the order value more the symmetries in them.

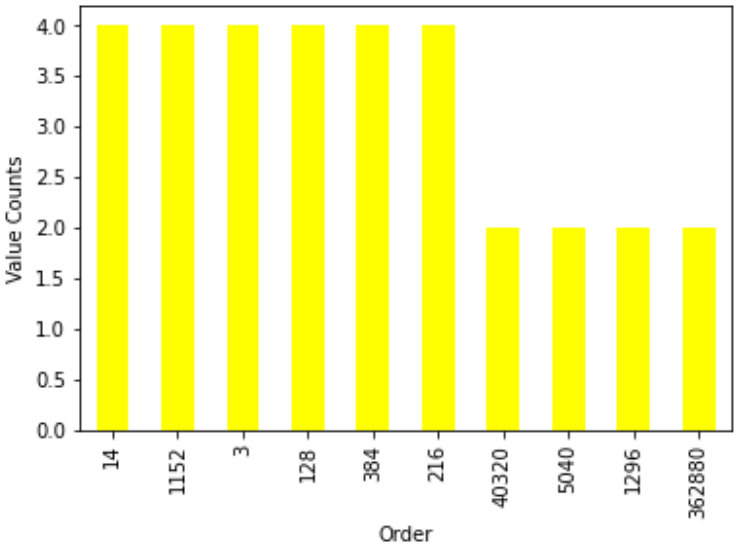

Figure 3.21: Bar graph of last 10 value counts in the column order

The orders with only 2 value counts (figure 3.21):

- 362880
- 1926
- 5040
- 40320

The highest order is 362880, but it has only two value count (figure 3.21). The null graph and fully connected graph have the order value of 362880 (figure 3.21). The graphs with small order numbers are asymmetrical. There are 45 distinct values, and the mean value is 6.34 (figure 3.22). All the graphs have one among the 45 values. The attribute consumes 2.1 MB of storage (figure 3.22).

The minimum value of the order is 1 (figure 3.23), and it belongs to the asymmetrical graphs.

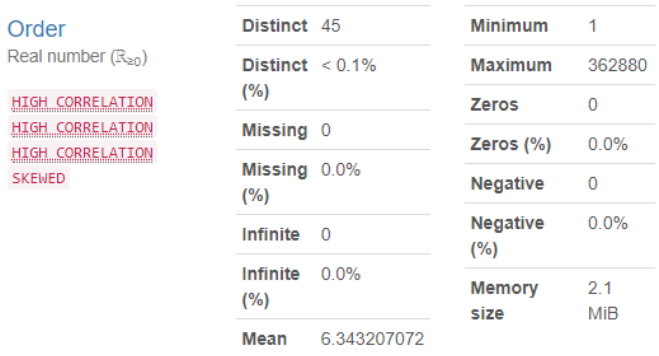

Figure 3.22: General statistics of the column order

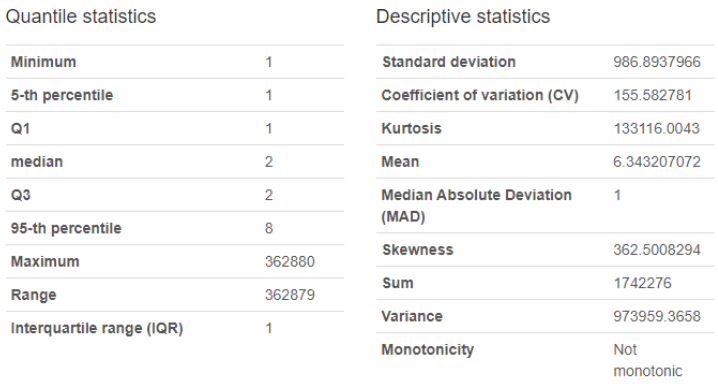

Figure 3.23: Quantile and Descriptive statistics of column order

The maximum value is 362880 (figure 3.23), which belongs to the null graph and fully connected graphs. Descriptive statistics is a method to analyse the features of the attributes. Q1 is the quantile 1, which is the 25% value of the attribute (figure 3.23). Around 50% of the data, value is 1, so the value of Q1 is 1 (figure 3.23). The Q3 is 75% of the data (figure 3.23).

$$Range = Maximum - Minimum$$

$$362880 - 1 = 362879$$

The value of the range is 362879 and the value of Q3 is 2 (figure 3.23). The standard deviation of the attribute is 986.893 (figure 3.23). The standard deviation measures the deviation rate of the values deviates from the mean [1]. The skewness is a measure to understand the data

distribution for sample data [16]. The skewness value is 362.5 (figure 3.23), and it is positive, so the attribute is positively skewed with a magnitude of 362.5. The variance is the square of the standard deviation. The variance of the data is 973959.36 (figure 3.23). The variance describes the distance of data from the mean. Most of the values are equal to 1, but still the variance is 973959.36. It is due to the presence of larger values like 362880, and 40320 and it is not monotonic. The total sum of all the values in the order is 1742276. The kurtosis value is 133116.043 and the mean absolute deviation value is 1(figure 3.23). The value of the Kurtosis suggests that the graph is sharp and it is leptokurtic. It shows high skewed and high correlation features.

Attribute State_Space_Orbit Analysis

The orbits are used to analyse the number of identical vertex in each orbit. If the number of orbits is equal to the number of vertices then there is no chance of building symmetry. There

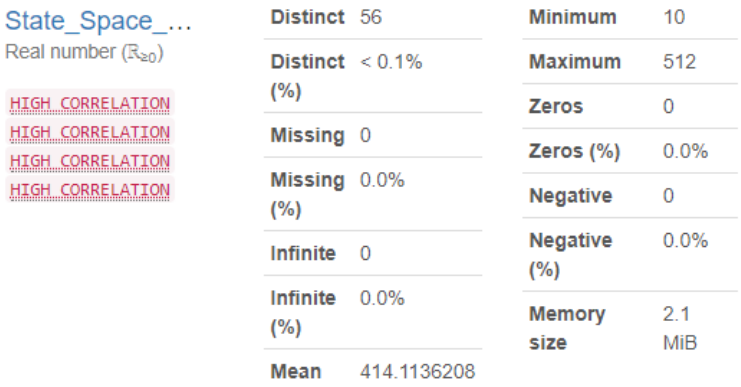

Figure 3.24: General statistics of the column state space orbit

are 56 distinct values in the state space orbits, and the maximum value of the state space orbit is 512, which is 2^9, with the percentage of distinct values being less than 0.1% (figure 3.24).

$$56/274668 = 0.002 < 0.1\%$$

The mean value of the state space order is 414.11 (figure 3.24), and the maximum value of the state space order is 512 (figure 3.24). If the value of the state space order for a graph is equal to 512 then the graph is symmetric. The memory consumed by the attribute state space order is 2.1MB (figure 3.24), there are no missing values in the data, and the percentage of zeros in the attribute is 0. State Space Orbits are highly correlated with Order (negative correlation) and three other fields (Asymmetry, Vertex Orbit Length, vertex orbit max (negative correlation)). The range of the data is $(10, 512)$.

$$State\ Space\ Orbits\ \alpha\ 1/Symmetry$$

State Space Order is indirectly proportional to symmetry. If the value of the state space orbit increases, then the number of symmetries will decrease (figure 3.25). The permutation of vertices on the null graph is also equal to 512, but the symmetrical graphs are reduced to avoid repetition. The state space orbit has a High correlation with four other attributes in the data.

$$State\ Space\ Orbits\ \alpha\ vertex\ orbit\ length \tag{3.1}$$

The vertex orbit length will be higher if the number of symmetries is lower. The vertex orbit

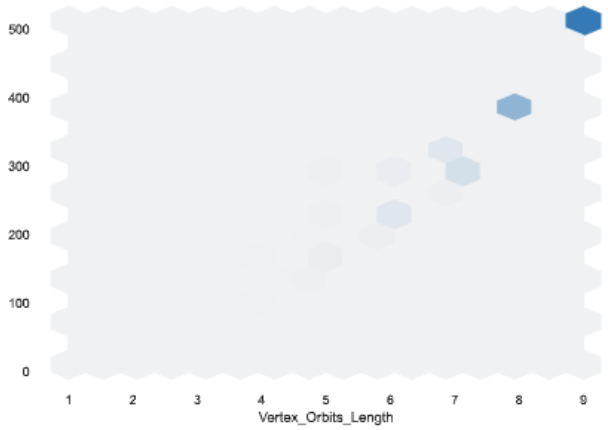

Figure 3.25: Interaction between vertex orbit length and the state space orbit

length is directly proportional to the symmetry, so state space orbit is directly proportional to vertex orbit length (figure 3.25).

$$State\ Space\ Orbits\ \alpha\ vertex\ orbit\ max \tag{3.2}$$

The attribute vertex orbit max will be higher if the attribute vertex orbit length is lower. The

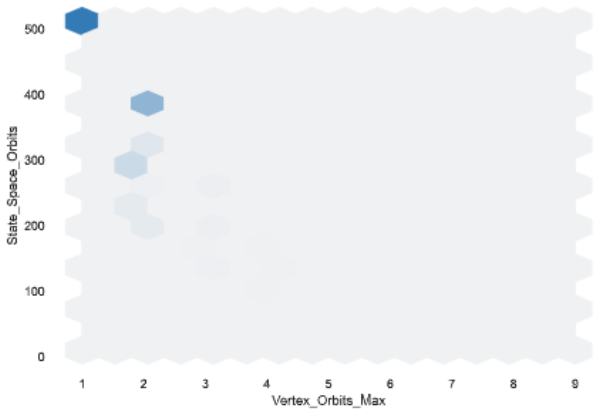

Figure 3.26: Interaction between vertex orbit max and the state space

vertex orbit max is inversely proportional to the vertex orbit length, so the state space orbit is inversely proportional vertex orbit max (figure 3.26).

Quantile statistics

Minimum	10
5-th percentile	216
Q1	320
median	384
Q3	512
95-th percentile	512
Maximum	512
Range	502
Interquartile range (IQR)	192

Descriptive statistics

Standard deviation	110.4217443
Coefficient of variation (CV)	0.2666460092
Kurtosis	-0.467365127
Mean	414.1136208
Median Absolute Deviation (MAD)	128
Skewness	-0.7531668473
Sum	113743760
Variance	12192.96162
Monotonicity	Not monotonic

Figure 3.27: Quantile and Descriptive statistics of state space orbit

From the figure 3.27, the value of Q1 is 320, and the value of Q3 is 512. The 5^{th} percentile value is 216, so the graphs with state space orbit less than 100 less (figure 3.27). The median or middle value of the attribute is 384. The range of attributes is 502, and the interquartile range is 192 (figure 3.27). The median is closer to the mean value, and the value count of all the state space orbits is taken in the data. The standard deviation of the data is 110.42, and the mean absolute deviation is 128 (figure 3.27). The standard deviation is less, so the distribution graph is not concentrated towards the centre. The attribute is not monotonic, and the kurtosis value is -0.467 (figure 3.27), which suggests that the peak is flatter and it is a playkutric graph. The graph is not much flatter as the magnitude of the kurtosis is 0.467. The sum of all the values is 113743760 (figure 3.27). The attribute is not monotonic. The variance is the square of the standard deviation, which is equal to 12192.96 (figure 3.27). The coefficient of the variation is 0.2666, and there are no negative values in the attribute.

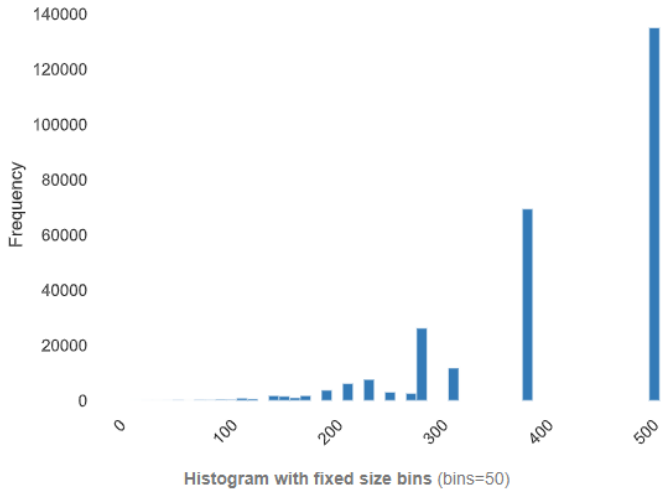

Figure 3.28: Histogram of state space orbit with its frequency

The histogram is the best method to interpret the numerical data in the attribute. The bins are of uniform interval in a frequency histogram. The data distribution can be misleading for the frequencies calculated in an unequal interval (figure 3.27).
There are 50 bins in the histogram, and the histogram is a normal distribution with negative skewness (figure 3.28). Most graphs have 512 state space orbits. The 49.2% of graphs, which

Value	Count	Frequency (%)
512	135004	49.2%
384	69382	25.3%
320	11772	4.3%
288	26194	9.5%
272	2602	0.9%
256	3034	1.1%
240	7640	2.8%
216	6178	2.2%
200	414	0.2%
192	3322	1.2%

Figure 3.29: Top 10 value count and frequency of the state space orbit

is 135004 graphs have their state space orbit value of 512 (figure 3.29). 69382 graphs have 384 state space orbit values which constitute 25.3% of the total value(figure 3.29). 74.5% of the graph's state space orbit's value is either 512 or 384. The rest of the 26.5% of graphs have 54 distinct values. The graphs with the top 5 state space orbits (figure 3.29):

- 512
- 384
- 288
- 320
- 240

The 91.2% of graphs have one of the top 5 state space orbit values. Astonishingly, the rest of the other graphs, which is 8.8% of graphs has 51 distinct state space orbit values.
The frequency of at least 10 values does not constitute 1% of the total (figure 3.30). There is only one graph, which has 26 state space orbits, and there is 4 state space orbits value (10, 18, 20 and 32), whose frequency is equal to 2 (figure 3.30).
The last 5 state space orbits values (figure 3.30):

- 26

- 10
- 18
- 20
- 32

Value	Count	Frequency (%)
10	2	< 0.1%
18	2	< 0.1%
20	2	< 0.1%
24	6	< 0.1%
26	1	< 0.1%
28	6	< 0.1%
30	14	< 0.1%
32	2	< 0.1%
36	8	< 0.1%
40	18	< 0.1%

Figure 3.30: Last 10 value count and frequency of the state space orbit

Attribute Vertex Orbit Length Analysis

The attribute vertex orbit length is the length of the list vertex orbits. If the vertex orbit is [1, 1, 6, 1], then the vertex orbit length is 4.
The vertex orbit length has 9 distinct values (figure 3.31). The values are in the range of 1 to 9 as the total number of vertices is equal to 9. The mean value is 8.04, and the memory consumed by the attribute is 2.1MB (figure 3.31). There are no null values or negative values, and all the values are real numbers in vertex orbit lengths value. The vertex orbit length has a high correlation with four other attributes.

$$Vertex\ Orbit\ Length\ \alpha\ 1/Symmetry$$

Vertex orbit length is inversely proportional to symmetry. The vertex orbit length of asymmetrical graphs is 9.

$$Vertex\ Orbit\ Length\ \alpha\ 1/Vertex\ Orbit\ Max$$

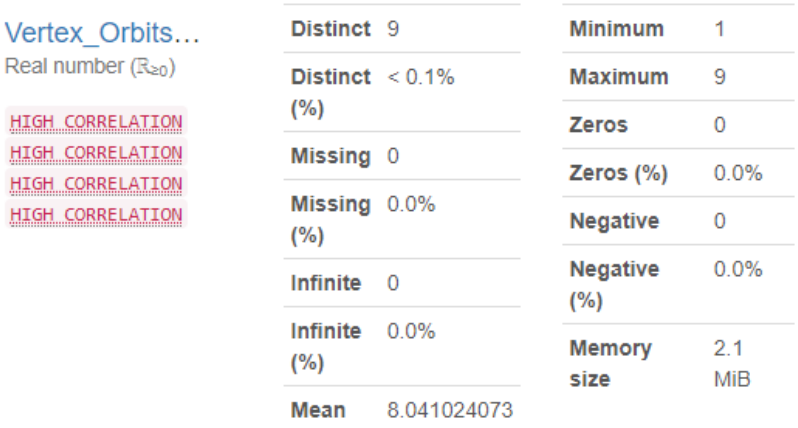

Figure 3.31: General statistics of vertex orbit length

Vertex orbit length is inversely proportional to the vertex orbit max. If the vertex max increases, then the number of values in the sublist will increase, which decreases the size of the list. Example of symmetrical graphs:

$$vertex = [\{1\}, \{2\}, \{3\}, \{4, 5, 6, 7\}, \{8, 9\}] = [1, 1, 1, 4, 2] \tag{3.3}$$

From the equation 3.3:

$$vertex\ orbit\ length = 5$$

$$[\{1\}, \{2\}, \{3\}, \{4, 5, 6, 7\}, \{8, 9\}] \longrightarrow list$$

$$\{4, 5, 6, 7\} \longrightarrow sub\ list$$

The length of the vertex orbit is 5. If there is more than 1 vertex in the sub-list of the graph, then the graph is symmetrical. The Q1 value or 25^{th} quartile value of the vertex orbit length is 7 (figure 3.32). The Q3 or 75^{th} quartile value is 9 (figure 3.32). These quartiles help to understand the skewness in the data using a box plot.

$$Range = 9 - 1 = 8$$

The interquartile is the range between the 25^{th} quartile and 75^{th}.

$$IQR = 9 - 7 = 2$$

The 5^{th} percentile value is 5 and the 95^{th} percentile value is 9 (figure 3.32). The standard deviation of the attribute is 1.22, and the mean is 8.04 (figure 3.32). The variance of the attribute is 1.5, and the coefficient of variation is 0.152 (figure 3.32). The median or $Q_0.5$ value is 8 (figure 3.32).
The skewness value is -1.37, it is negatively skewed or right skewed, which interprets the mean

Quantile statistics

Minimum	1
5-th percentile	5
Q1	7
median	8
Q3	9
95-th percentile	9
Maximum	9
Range	8
Interquartile range (IQR)	2

Descriptive statistics

Standard deviation	1.228536955
Coefficient of variation (CV)	0.1527836435
Kurtosis	1.488118715
Mean	8.041024073
Median Absolute Deviation (MAD)	1
Skewness	-1.372260748
Sum	2208612
Variance	1.50930305
Monotonicity	Not monotonic

Figure 3.32: Quantile and Descriptive statistics of vertex orbit length

as greater than the median, but the difference between the mean and median is a negligible value (figure 3.32). The kurtosis value is 1.48, which is positive but less than 3 (figure 3.32) [16]. The distribution is platykurtic, but it is closer to three, so it is closer to becoming a normal distribution (figure 3.32). The sum of all the values is equal to 2208612 (figure 3.32).

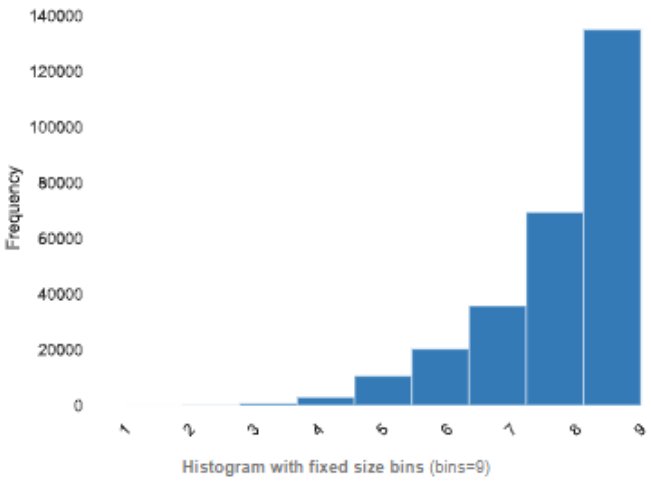

Figure 3.33: Histogram of vertex orbit length with its frequency

It is easy to interpret the values like skewness and kurtosis in a histogram. The histogram is right skewed or negative skewed. The total number of bins is 9 (figure 3.33). If the length of the vertex orbit increases, the number of symmetries decreases. The number of graphs increases with the increase in the length of the vertex orbits. The size of all the bins are equal in the interval.

49.2% of the graphs have the vertex orbit length as 9, and the graphs are asymmetrical as the

Value	Count	Frequency (%)
9	135004	49.2%
8	69382	25.3%
7	35752	13.0%
6	20266	7.4%
5	10548	3.8%
4	2920	1.1%
3	657	0.2%
2	130	< 0.1%
1	9	< 0.1%

Figure 3.34: The value count and frequency of vertex orbit length

vertex orbit length is equal to 9 (figure 3.34). Around 75% of the graphs has either 9 or 8 as the vertex orbits length (figure 3.34). Percentage of graphs having vertex orbit lengths less than five is around 5% (figure 3.34).

Attribute Vertex Orbit Max Analysis

The attribute vertex orbit max helps to classify the graphs into symmetrical and asymmetrical. The vertex orbit max is also helpful in analysing the state of the vertex orbits and their permutations. The type of the data is numerical (integer).

Vertex_Orbits…
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	9	**Minimum**	1
Distinct (%)	< 0.1%	**Maximum**	9
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1.578440153	**Memory size**	2.1 MiB

Figure 3.35: General Statistics of vertex orbit max

There are 9 distinct values in attribute vertex orbit max and the distinct percentage is less than

0.1% (figure 3.35). The minimum value is 1, and the maximum value is 9 (figure 3.35).
Example of Vertex Orbit Max:

$$vertex = [\{1\}, \{2\}, \{3\}, \{4\}, \{5, 6, 7\}, \{8, 9\}] = [1, 1, 1, 1, 3, 2] \tag{3.4}$$

From the equation 3.4:

$$vertex\ orbit\ max = 3$$

The maximum value among the length of the sublists is the vertex orbit max. There is no possibility of having a null set, so there are no zeros in the list. The memory consumed by the attribute is 2.1MB (figure 3.35) . The vertex orbit max is having high correlation with four other attributes (figure 3.35).

$$Vertex\ Orbit\ Max\ \alpha\ Symmetry$$

The vertex orbit max value increases with an increase in symmetry. The null graph has the highest symmetry, and the vertex orbit max value is 9.

$$vertex = [\{1, 2, 3, 4, 5, 6, 7, 8, 9\}] = [9] \tag{3.5}$$

From the equation 3.5:

$$vertex\ orbit\ max = 9$$

The asymmetrical graph's vertex orbit max value is 1.

$$vertex = [\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}, \{9\}] = [1, 1, 1, 1, 1, 1, 1, 1, 1] \tag{3.6}$$

From the equation 3.6:

$$vertex\ orbit\ max = 1$$

The asymmetrical graphs will have the minimum value, and the graph with the highest number of symmetries will have a high value. The Q1 value or 25^{th} quartile value of the vertex orbit max is 1 (figure 3.36). The Q3 or 75^{th} quartile value is 2 (figure 3.36). These quartiles help to understand the skewness in the data using a box plot.

$$Range = 9 - 1 = 8$$

The interquartile is the range between the 25^{th} quartile and 75^{th}.

$$IQR = 2 - 1 = 1$$

The 5^{th} percentile value is 1 and the 95^{th} percentile value is 2 (figure 3.36). The interquartile range and all quartiles are the quantitative values to illustrate a box plot. The standard deviation of the attribute is 0.6599, and the mean is 1.57 (figure 3.36). The variance of the attribute is

Quantile statistics

Minimum	1
5-th percentile	1
Q1	1
median	2
Q3	2
95-th percentile	2
Maximum	9
Range	8
Interquartile range (IQR)	1

Descriptive statistics

Standard deviation	0.659972588
Coefficient of variation (CV)	0.4181169534
Kurtosis	4.895260469
Mean	1.578440153
Median Absolute Deviation (MAD)	1
Skewness	1.41946383
Sum	433547
Variance	0.4355638169
Monotonicity	Not monotonic

Figure 3.36: Quantile and Descriptive statistics of vertex orbit max

0.4355, and the coefficient of variation is 0.418 (figure 3.36). The median or $Q_0.5$ value is 2 (figure 3.36).

The skewness value is 1.419 (figure 3.36), and it is positively skewed or known as left skewed. The mean is less than the median, but the difference between the mean and median is a negligible value. The kurtosis value is 4.89, which is positive and greater than 3, it is leptokurtic, but it is closer to three, so it is closer to becoming a normal distribution [16]. The sum of all the values is equal to 433547. It is not a monotonic function (figure 3.36).

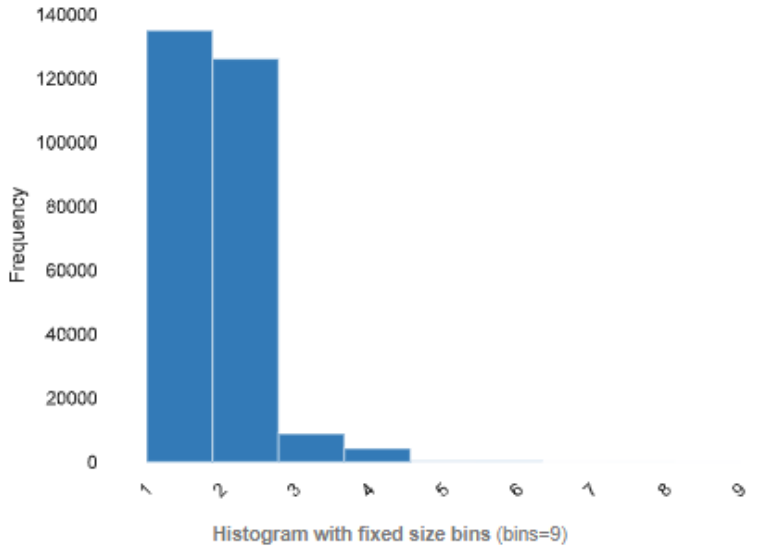

Figure 3.37: Histogram of vertex orbit max

It is easy to interpret the statistical details like the skewness and kurtosis in the histogram, and this histogram is right skewed or negative skewed (figure 3.37). The total number of bins is 9 (figure 3.37). If the maximum length of the sublist of the vertex orbit increases, the number of

symmetries increases. The number of graphs decreases with the increase in the attribute vertex orbits max (figure 3.37). The size of all the bins is equal in the interval.

$$vertex = [\{1\}, \{2\}, \{3\}, \{4\}, \{5, 6\}, \{7\}, \{8\}, \{9\}] = [1, 1, 1, 1, 2, 1, 1, 1] \tag{3.7}$$

From the equation 3.7:

$$vertex\ orbit\ max\ 1 = 2$$

$$vertex = [\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7, 8\}, \{9\}] = [1, 1, 1, 1, 1, 2, 1, 1] \tag{3.8}$$

From the equation 3.8:

$$vertex\ orbit\ max\ 2 = 2$$

Though the vertex orbit max 1 and 2 are equal, but they are not symmetries of each other (figure 3.38).

Value	Count	Frequency (%)
1	135004	49.2%
2	126104	45.9%
3	8820	3.2%
4	4217	1.5%
5	230	0.1%
6	238	0.1%
7	20	< 0.1%
8	26	< 0.1%
9	9	< 0.1%

Figure 3.38: The value and frequency count of vertex orbit max

The 49.2% of the graph falls under the value 1, which is 135004 graphs have 1 as the maximum length of their sublists (figure 3.38). 45.9% graphs have 2 as their maximum length among the sub-lists (figure 3.38). 94.1% of the graphs have either the maximum length of sublist as either 1 or 2. The other 5.9% has 6 distinct vertex orbit lengths. If the vertex orbit max value is equal to 1, then the graphs are asymmetric.

Attribute Asymmetry Analysis

The attribute asymmetry is a Boolean data type. The boolean data type is binary data in the form of words. The data can either be true or false. The attribute asymmetry can be created by vertex orbit length.

If Vertex orbit length is equal to the number of vertices, then the graph is assumed to be asymmetric.
Asymmetry:

$$vertex\ orbit\ length = number\ of\ vertices$$

$$X = \begin{cases} True & \text{if, vertex orbit length = 9} \\ False & \text{otherwise} \end{cases} \tag{3.9}$$

From the equation 3.9, if the vertex orbit length is 9, then the code returns True. If it is not equal to 9, it returns false.

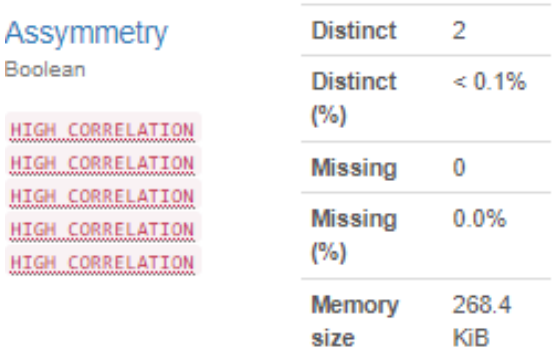

Figure 3.39: Description of the column asymmetry

Asymmetry has a high correlation with four other attributes (figure 3.39). The correlation will be the same as the vertex orbit length. The asymmetry memory size is 268.4KB (figure 3.39). The

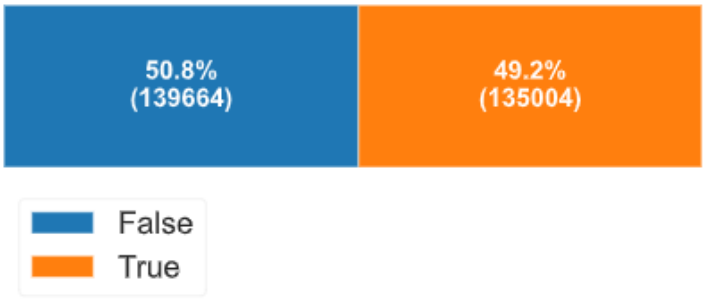

Figure 3.40: categorical plot of the asymmetry column

blue colour represents symmetric graphs, and the orange represents asymmetric graphs (figure 3.40). The 50.8% of the graphs, which is 139664 graphs in the data are symmetric (figure 3.40). The 49.2% of the graphs, which is 135004 graphs in the data are asymmetric (figure 3.40). The 50.8% graphs' computational storage can be reduced (figure 3.40). The percentage of symmetric graphs for larger analysis will reduce the computational memory.

Network Analysis using NetworkX Library

The networkx is a python library used specifically to analyse the dynamics, structure and creation of networks or graphs. The library allows importing graph files through the command `read_graph6` (figure 3.41). The command directly allows python to import the graphical file without processing them in the GAP software.

```
import networkx as nx
g1 = nx.read_graph6 ("graph9.g6")
```

Figure 3.41: code for importing graph6 file by using networkx library

The networks can be created using the command `networkx.draw\_circular()` (figure 3.42) [6]. The network is created to illustrate and validate the data obtained from the GAP software.

```
nx.draw_circular(g1[100000])
plt.show()
```

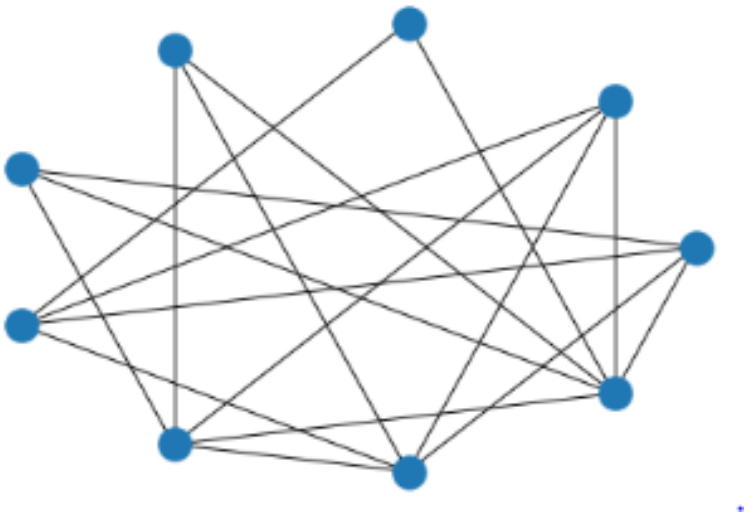

Figure 3.42: code and the network created by using networkx library

The network is drawn in a circular format (figure 3.42). The graphs can be drawn in random, shell, spectral, spring and circular. The possible structure in networkx library:

- Random
- Shell
- Spectral

- Spring
- Circular

In the automorphism, shape of the graph is not important, but the structure of the graph is important. The statistical analysis helps to identify the possible symmetries and calculate the probability of the situation. Example the probability of the next vertex being infected. The statistical details of automorphism group is a primary evidence for concluding any assumptions taken.

Chapter 4

Results

The objective of our model is to analyse the symmetries for the networks with given vertices. The symmetries are analysed by plotting logarithmic points between the order and state space. The logarithmic values are taken to analyse the trend or relation with accuracy.

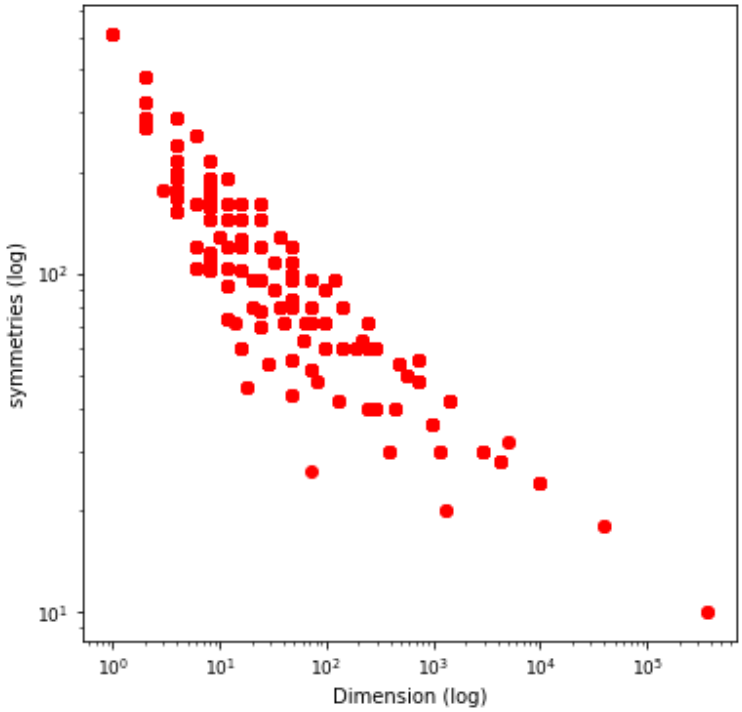

Figure 4.1: Dimension versus Symmetry Graph

The graphs show negative linearity, which means the number of symmetries decreases with an increase in the dimension. The symmetry of the graph decreases with an increase in the order or dimension (figure 4.1). From the figure 3.22 there are only 45 distinct values. The graph contains 45 points, which are depicted in the figure 4.1 and the rest of the graphs are at the same point. There are around 135000 graphs on the last point, they are exactly on the same point, and

this last point represents the asymmetric graphs.

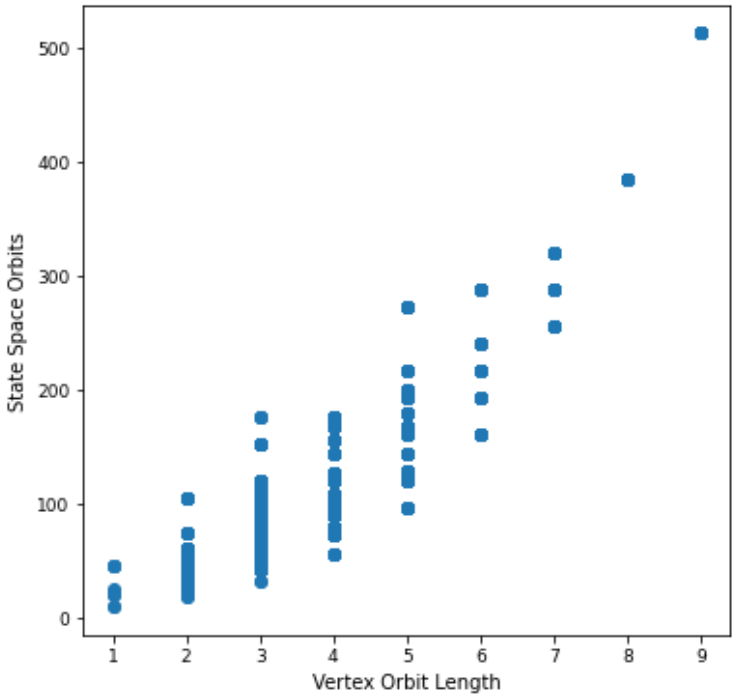

Figure 4.2: Vertex Orbit Length vs State Space Orbit

Vertex Orbit Length 1:

$$[1, 2, 1, 2, 3] = 5 \tag{4.1}$$

Vertex Orbit Length 2:

$$[1, 2, 2, 2, 2] = 5 \tag{4.2}$$

Vertex Orbit Length 3:

$$[1, 1, 1, 3, 3] = 5 \tag{4.3}$$

Vertex Orbit Length 4:

$$[1, 4, 2, 1, 1] = 5 \tag{4.4}$$

Vertex Orbit Length 5:

$$[1, 5, 1, 1, 1] = 5 \tag{4.5}$$

Vertex Orbit Length 6:

$$[1, 3, 2, 2, 1] = 5 \tag{4.6}$$

There are 6 different types of vertex orbit with their length is equal to 5, but their vertex orbits are different (equations 4.1, 4.2, 4.3, 4.4, 4.5, 4.6). The state space of the 6 vertex orbit differs, so there are multiple points in a single vertex. These multiple points are distinct values, and they are viewed in the graph, but some values have the same state space orbits, and they are exactly on the same point, and it is hard to distinguish them. In vertex orbit length 9 the state space orbit is 512, and it same for all, so there are no multiple state space orbit values.
From the equation 3.1, the vertex orbit length is directly proportional to the state space orbits. The number of symmetries decreases with the increase in the vertex orbit length. The chart (figure 4.2) shows a steady increase proving the equation 3.1 is true. All the state space orbit values of vertex orbit length 9 are 512, but the other for vertex orbit lengths has multiple values.

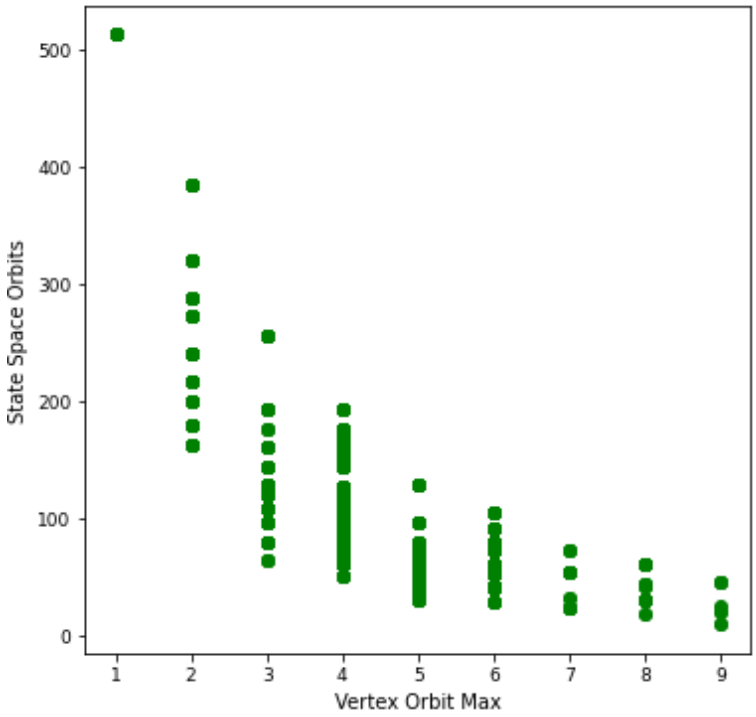

Figure 4.3: Vertex Orbit Max vs State space Orbit

Example in vertex orbit Max 2:
Vertex Orbit Max 1:

$$[1, 2, 1, 2, 1, 1, 1] = 2 \tag{4.7}$$

Vertex Orbit Max 2:

$$[1, 2, 2, 2, 2] = 2 \tag{4.8}$$

Vertex Orbit Max 3:

$$[1, 1, 1, 2, 2, 2] = 2 \tag{4.9}$$

Vertex Orbit Max 4:

$$[1, 1, 1, 1, 1, 2, 1, 1] = 2 \tag{4.10}$$

There are 4 different types of vertex orbit with their maximum length of their sublist is equal to 4, but their vertex orbits are different (equations 4.7, 4.8, 4.9, 4.10). The state space of the four vertex orbit differs, so there are multiple points in a single vertex. These multiple points are distinct values, and they are viewed in the graph, but some values have the same state space orbits, and they are exactly on the same point, and it is hard to distinguish them. In vertex orbit max 1 the state space orbit is 512, and it same for all, so there are no multiple state space orbit values.

From the equation 3.2, the vertex orbit length is inversely proportional to the state space orbits. The number of symmetries increases with the increase in the vertex orbit Max. The 50% of the data has symmetries in them, so if the dimension of the graph increases the symmetries decreases.

The symmetries increases with the increase in the number of vertices, and the process for analysing symmetries is long. In other words, the number of asymmetrical graphs decreases with an increase in the number. Though the large vertex graphs are exploited with the method of approximation. The data for the 12 vertices graph is 10 billion, and if the 50% of the graph is reduced by symmetry, still the number of graphs used for analysis is 5 billion, which is still a huge number. The symmetries and approximation are methods to analyse the large dataset.

Chapter 5

Conclusion

The dynamic process of the network provides information about its state. The network's dynamic process revolves around the concept of Markov's Chain. The permutation of the network with large vertices generates a huge number of graphs, which requires lots of computational storage. These graphs are analysed by reducing the size by the method of symmetry. Symmetries are the different combinations of the network without affecting its structure [18]. A pipeline is a sequential process including data extraction, symmetry analysis and statistical data analysis built to analyse the symmetries. The mathematical concepts of graph and group theory form the core idea for analysing the network's symmetries. Chapter 2 explains the mathematical concepts required for the analysis. The data was extracted from the Brendon Mckay website in graph format. The digraph package in GAP software was used to analyse the data in the graph format and converts the output into CSV format. The python carried out all the required statistical analyses of the CSV file due to its versatile library collection. The results from the statistical analysis help to understand the percentage of symmetries in a combination of a network. The process requires high storage, which was solved by the usage of HPC (High-Performance Computer). The pipeline produces the required result to calculate the state of the data. The applications of the dynamic process on networks in real-time include influence on voting behaviour, checking the virus transmission through the internet and the spread of disease through the epidemic.

Bibliography

[1] math.net . Standard deviation.

[2] Tutorial Point . Graph theory - types of graphs, 2015.

[3] M Aslam and Akbar Ali. The graph δ 2n–1 is an induced subgraph of a johnson graph. *Int. J. Contemp. Math. Sciences*, 7:369–376, 2012.

[4] Sayantani Biswas. Omicron variant symptoms: 5 things you need to know, 11 2021.

[5] Jan De Beule, Julius Jonušas, James Mitchell, Michael Torpey, Wilf Wilson, Stuart Burrell, Reinis Cirpons, Luke Elliott, Max Horn, Christopher Jefferson, Markus Pfeiffer, Chris Finn, and Smith White. Digraphs, 2014.

[6] networkx developers. Drawing — networkx 2.8.6 documentation.

[7] Sergey Finashin. Graph theory, part 1 lecture notes in math 212 discrete mathematics, 2020.

[8] guptavivek. Complement of graph, 02 2022.

[9] Sarada Herke. Graph theory faqs: 02. graph automorphisms.

[10] Brendon Mckay. Combinatorial data.

[11] Wes Mckinney. General functions pandas 1.0.3 documentation.

[12] Abbe Mowshowitz, Matthias Dehmer, and Frank Emmert-Streib. A note on graphs with prescribed orbit structure. *Entropy*, 21:1118, 11 2019.

[13] Sebastian Payne. Rishi sunak to face liz truss in battle to become uk prime minister. *Financial Times*, 07 2022.

[14] H.N.de Ridder. List of small graphs, 02 2022.

[15] Joseph Rocca. A brief introduction to markov chains, 02 2019.

[16] Sakshi. Kurtosis and skewness - machine learning concepts, 02 2022.

[17] Manikanta satyala. Graph theory in network system, 04 2018.

[18] Jonathan A. Ward. Dimension-reduction of dynamics on real-world networks with symmetry. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 477:20210026, 07 2021.

[19] Jonathan A. Ward and John Evans. A general model of dynamics on networks with graph automorphism lumping. *Studies in Computational Intelligence*, pages 445–456, 12 2018.

[20] Wikipedia. Network topology, 11 2021.