Department of Physics and Astronomy

University of Heidelberg

Master thesis

in Physics

submitted by

Julian Heiss

born in Homburg

2018

Predicting Atomic Transitions

Through the Application of Network Theory

to Spectroscopic Data

This Master thesis has been carried out by Julian Heiss

at the

Physical Institute Heidelberg

under the supervision of

Prof. Dr. Matthias Weidemüller

Vorhersage von atomaren Übergängen durch den Einsatz der Netzwerktheorie für spektroskopische Daten:

Diese Arbeit legt einen datengetriebenen Ansatz zur Betrachtung atomarer Spektraldaten dar. Wir bilden Spektraldaten auf ein Netzwerk ab, in welchem Knoten Energiezustände repräsentieren und eine Verbindung zwischen Knoten existiert, wenn zwischen den dazu entsprechenden Zuständen ein atomarer Übergang gemessen wurde. Wir untersuchen die gängigsten Netzwerkeigenschaften dieser sogennanten spektroskopischen Netzwerke oder vergleichen sie mit denen von Netzwerken, die mit Zufallsgraphen-Modellen und skalenfreien Netzwerkmodellen erzeugt wurden. Wir zeigen, dass die Netzwerkaufteilung spektroskopischer Netzwerke in Gruppen von Knoten die zugrundeliegenden Symmetrien der atomaren Systeme widerspiegelt. Außerdem legen wir dar, wie durch die Bestimmung fehlender Verbindungen in den Netzwerken ungemessene atomare Übergänge ohne Zuhilfenahme eines mikroskopischen Atommodells vorhergesagt werden können. Wir prüfen und vergleichen die Vorhersagekraft von vier ähnlichkeitsbasierten Methoden, zwei Maximum-Likelihood-Methoden und einer Perturbationsmethode in den spektroskopischen Netzwerken von Wassterstoff, Helium, Kohlenstoff und Eisen. Die nicht ähnlichkeitsbasierten Methoden liefern die zuverlässigsten Prognosen, wobei die Perturbationsmethode die besten Ergebnisse aufweist.

Predicting Atomic Transitions Through the Application of Network Theory to Spectroscopic Data:

This thesis demonstrates a data-driven approach to treating atomic spectroscopic data. We map spectral data onto a network in which nodes represent energy states and a link exists between two nodes if an atomic transition has been observed between the corresponding states. We perform an analysis of the most common network properties of these so-called spectroscopic networks and compare them with those of networks generated from random graph and scale-free models. We show that the node community structure of spectroscopic networks reflects the underlying symmetries of atomic systems. In addition, by inferring missing links of the networks, we demonstrate that unmeasured atomic transitions can be predicted without having to construct a microscopic model of the atom. We test and compare four similarity-based methods, two maximum likelihood methods and a perturbation-based link prediction method on spectroscopic networks of hydrogen, helium, carbon and iron in terms of their predictive power. The non-similarity-based methods yield the most reliable results, with the structural perturbation method performing best.

Contents

1 Atomic Spectra and Network Science: A New Approach

It is an empirical fact that most natural and engineered systems are composed of subsystems that exhibit some kind of connectivity and dependencies [17]. Network science tries to describe this connectivity to gain insights in these systems. In its simplest form, a network is a set of items, usually called nodes, with connections between them, called links [34].

Leonhard Euler's solution of the Königsberg bridge problem is often celebrated as the first proof in the field of network theory, which in its mathematical form is also called "graph theory" [34]. Since then, there have been advancements in this mathematical field, but the advent of computers and their ability to efficiently gather, store and process data on far larger scales than ever before particularly spurred the interest in network research over the last two decades [34]. Furthermore, a reason for this certainly is the broad applicability of network science to various and diverse scientific fields, ranging from the study of social interactions to food webs and protein-protein interactions [35].

Network science has also attracted attention in the physics community which lead to an exchange between the fields. Notions from statistical physics were for example used to describe the evolution of networks [4]. However, there are only few examples where fundamental physical systems have been modelled in terms of networks. Halu *et al.* studied the phase transitions of the Bose-Hubbard model on networks [19], and Kulvelis *et al.* considered quantum transport processes on networks [27]. Yet in both cases only special topologies, such as tree-like networks [27], were considered. A more recent use of network theory for physical systems is the work by Valdez *et al.* in [44], where one-dimensional cases of both the Ising and the Bose-Hubbard model are mapped onto fully connected networks in which nodes represent particle

sites and link weights are determined by the quantum mutual information of each particle pair. They demonstrated that in this representation simple network measures serve as order parameters of these models. We see that in contrast to the other fields, where networks are used for the investigation of empirical data, in the field of physics networks have so far been primarily considered in heavily idealised cases.

In contrast, in this project we study if concepts from network theory prove to be meaningful when applied to real-world physical data sets. For this, we choose to represent spectral data as a network. There is an intuitive way to represent this system in terms of a network: we use the energy eigenstates as nodes and connect two nodes by a link if a transition between those states has been observed in experiments. This transition from physical quantities to abstract network entities is depicted in Figure 1.1.
Most importantly, since the behaviour of such systems is well described by the quantum mechanical theory of atoms, we have a microscopic model of the network that enables us to validate our findings. This is not the case for empirical data in other scientific fields and puts us in a unique position.

We want to find out if the mapping of this spectral data onto a network gives rise to the description of non-trivial physical notions in terms of network properties. How do the underlying symmetries of the physical system manifest themselves in the network structure? Can this data-driven approach provide insights that go beyond those of the microscopic model for atoms?

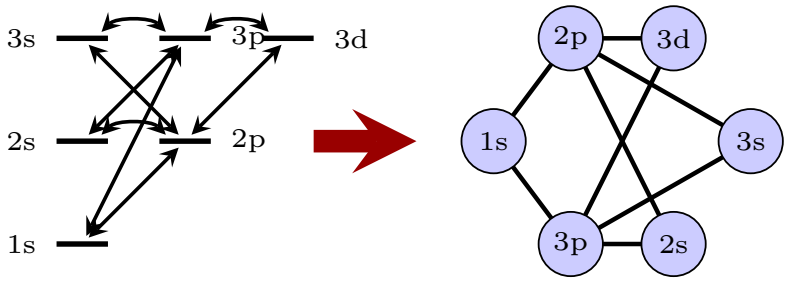

Figure 1.1: Energy levels and electric dipole transitions of hydrogen up to $n = 3$ as a Grotrian diagram (left). We map this system onto a network by representing energy states as nodes and the transitions between them as links (right). Figure provided by A. Kekić.

With the exception of the work of Cszászár and Furtenbacher *et al.* in [8] and [15], we know of no scientific work in which spectral data has been represented as a network. In said work though, the authors investigated molecular systems, whereas we are considering atomic systems.

The outline of this work is as follows: first, we introduce the mapping of data onto networks in detail and study if basic statistical properties that characterise the structure and behaviour of the networks relate to the physical properties of the atomic system. In a second step, we seek to capture the network structure by partitioning the constituents of the network into groups. We then study if there is a physical origin to the grouping patterns that appear. Lastly, we aim to infer missing links in the network on basis of the structural network data, enabling us to predict atomic transitions solely on the basis of data, *i.e.* without the use of a physical model of the atom.

2 Spectroscopic Networks

Network science evolved from the study of mathematical models, but is nowadays mostly used as a framework to investigate systems that are based on observational data [2]. Newman states in his book [35] that such real-world networks can be divided roughly into different classes, such as social networks or technological networks, that have distinct properties and therefore networks within a class are often treated with similar techniques.

We will use the methods of network science as means to learn about the physical system. Hence throughout this thesis we will interpret and compare the properties of the studied networks in the light of knowledge that we possess due the the quantum mechanical theory of atoms.

The discussion in this chapter is guided by two current textbooks on network science: *"Networks: an introduction"* by Newman [35] and *"Network science"* by Barabási [2]. In particular, we seek to answer the following questions in this chapter:

- How does the explicit mapping of spectral data onto networks work?
- Can we learn anything about these networks themselves or the physical system they are representing, by looking at the most common network properties and characteristics?
- Can these networks be attributed to any existing class of network models?

Concepts of network theory and atomic physics that are central to the discussion will not be explained up front, but rather introduced along the way as needed.

2.1 Selection of Data

The work in this master thesis will be carried out on the back of five representative networks of four different atoms which are listed in Table 2.1: hydrogen, helium, carbon and iron. With the exception of one network, all were constructed from data that we obtained from the *Atomic Spectra Database* (ASD) of the *American National Institute of Standards and Technology* (NIST) [25]. The remaining network is based on calculations of Jitrik and Bunge in [22] and is the second network describing the hydrogen atom.

The four atoms have been chosen for several reasons: Hydrogen and helium are theoretically well investigated, as they are the simplest atomic systems and therefore the atomic workhorses of quantum mechanics. Carbon as a six-electron-system and iron, with its 26 electrons, are systems with increasing complexity such that it is not possible any more to find exact solutions of the Schrödinger equation for iron. If we were to gain insights in this system that can not be obtained by the common current quantum mechanical methods, it would be a valuable contribution. Also, because iron has the highest nuclear binding energy, this element is of interest in fusion processes and of major importance in astrophysics [1][10].

The data sets that we are using are tables where each row contains the features corresponding to a spectral line, such as configuration of the lower and upper level and Ritz wavelength. A full description of the available features in this data set can be found in [37].

We will be using a simple graph representation throughout this work, *i.e.* an unweighted, undirected network without self-loops or multiple connections between two nodes. Atomic transitions between two states are represented as *links*. Further we only use the fact that the lower and upper levels of a transition constitute the end points of the corresponding link - its *nodes*. By doing this for each transition in the data set we construct a network that are going to call a *spectroscopic network*.

The energy levels are uniquely identified by a identification number also included in the data set which we use to prevent representing the same energy level by two distinct nodes in the network. The identification number will hence only be used

when creating the network or when evaluating results. The basis for any following work is the structural data of the network, the unique way the energy levels are connected - where by connection we mean that a change of an electron from one level to the other via emission or absorption of a photon is permitted by the laws of nature.

In the creation of networks from the raw data from the Atomic Spectra Database we omitted energy levels that were calculated solely by means of the Rydberg formula. Also, we only consider the largest component of the networks, *i.e.* the connected subgraph with the highest number of nodes. For the theoretically obtained hydrogen network H_{JB} there are no such filters in place, yet the authors only considered following states in their data tables: for $l \leqslant 4$ all states up to $n < 7$ are included, where l refers to the azimuthal quantum number and n to the principal quantum number of the energy state. For $l \geqslant 5$ all states with $n < l + 3$ and $n < 26$ are included. Additionally up to $n = 25$ all states with $l \leqslant 3$ are included. For all atomic transitions up to the order of electric and magnetic octupoles between these states the authors derived the respective wavelength, transition probability and further physical quantities. This selection of energy states is depicted in Figure 2.1. Based on this data we constructed a network consisting of 289 nodes and 15 793 edges which is in the following also referred to as H_{JB} (named after the surnames of the authors).

All of the further discussions will be based on the five networks listed in Table 2.1, if not stated otherwise. This table also contains several basic network properties that will be discussed in the following sections.

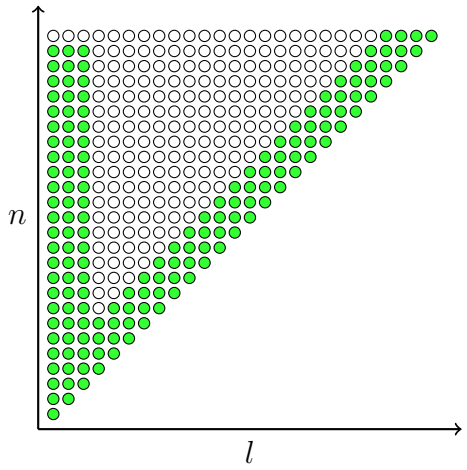

Figure 2.1: Energy states of hydrogen included in the data set provided by Jitrik and Bunge [22]. Each dot represents a hydrogen state and its position encodes its principal quantum number n and its azimuthal quantum number l, starting with the ground state at the origin. States marked in green and all E1, E2, E3, M1, M2 and M3 transitions between them are included in the data set. This data set is the basis for the network H_{JB}. Figure provided by A. Kekić.

Table 2.1: The basic structural features of the five studied example networks. N and L are the total number of nodes and links, respectively. D denotes the network density. r is the degree assortativity coefficient and C the transitivity coefficient of the network (which is sometimes also referred to as the clustering coefficient, see text). L_{E1}/L is the fraction of links in the network for which the transition they represent is of the electric dipole type (E1). The column *network* lists the designation by which we will sometimes refer to the respective network. *Source* indicates the data base from which we obtained the spectral data.

Atom	Network	Source	N	L	D	r	C	L_{E1}/L
Hydrogen (theoretical)	$\mathrm{H_{JB}}$	[22]	289	15 793	0.35	0.90	0.97	0.31
Hydrogen (experimental)	$\mathrm{H_{NIST}}$	[25]	66	278	0.11	−0.31	0.24	0.79
Helium	He	[25]	191	2 299	0.12	0.43	0.021	0.98
Carbon	C	[25]	180	1 377	0.075	−0.40	0.012	0.996
Iron	Fe	[25]	846	9 897	0.025	−0.29	0.029	0.989

2.2 Standard Analysis

In this section we will discuss several measures and metrics that describe the network structure and are commonly used in the standard literature about the study of networks. This discussion aims to give some insights in what properties are characteristic for spectroscopic networks. This list of properties considered in this discussion is far from exhaustive, but as we study the properties of spectroscopic networks, we are going to notice that the questions that these basic properties are often not able to capture the full complexity of atomic systems, and we should thus turn to more complex methods.

Furthermore we will compare the two data sets for hydrogen - the experimental data in the *Atomic Spectra Database* of the NIST [25] and the theoretical data by Jitrik and Bunge in [22] - side-by-side with respect to their network properties and their inherent structure. The main difference of these data sets is that the NIST data has been obtained by experimental observation, whereas the theoretical data was generated from first principles quantum mechanical calculations. We are going to see how this difference in origin affects the structural properties of the spectroscopic networks.

Ultimately, this discussion of network structure is not an end in itself, but each pattern or anomaly on the level of networks should be understood as well in the language of quantum systems.

2.2.1 Basic Properties

Network Density A simple graph only consists of two entities: nodes and links. The abundance of links is limited by the number of nodes. One of the most basic network properties is therefore the *network density*, as it answers the obvious question of how close we are to this limit: the *density* D of a network is the ratio of edges that are present L to the maximum possible number of edges $\binom{N}{2}$, where N is the number of nodes in the network. It follows [35, 6.9]

$$D = \frac{L}{\binom{N}{2}} = \frac{2L}{N(N-1)}\,. \tag{2.1}$$

As listed in Table 2.1, of all considered spectroscopic networks the theoretically obtained network H_{JB} is the most dense network with $D_{H_{JB}} = 0.35$. The network with the second highest density is helium with $D_{He} = 0.12$, followed by H_{NIST} with a vlaue of 0.11. We trace the higher density of H_{JB} back to the fact, that is this data set also high-order transitions were included. This was explained in Section 2.1. This means that the other networks should primarily consist of dipole lines. As a matter of fact, as listed in Table 2.1 the fraction of edges based on electric dipole transitions $\frac{L_{E1}}{L}$ in the network H_{NIST} is close to 80%. The other experimental networks even have rates of 98% and higher. In contrast, the line types in H_{JB} are more diverse and the fraction of electric dipole transitions is around 30%.

These numbers show that the electric dipole line type is predominant in the networks generated from the NIST data and constitutes the backbone of the spectroscopic networks. Because of *Laporte's rule* that prohibits transitions between states of same parity [5], the electric dipole transition create a bipartite structure in the network, *i.e.* the nodes are split into two sets according to the parity of the states. There are exceptions to the selection rules that weaken this bipartite structure: *e.g.* due to failure of the LS-coupling, the spin rule $\Delta S = 0$ can be violated, leading to semi-forbidden electric dipole transitions; magnetic dipole transitions will occur only between states belonging to the same fine structure multiplet; and electric quadrupole transitions must happen between states of the same parity (in contrast to electric dipole transitions) [5]. Although the transition probability of these higher-order transitions types are typically negligibly small compared to the probabilities of electric dipole lines, in the simple graph representation that we chose, this difference is disregarded. Thus, the backbone of the spectroscopic network is a bipartite graph, enriched by transitions that violate the set of selection rules for electric dipole transitions.

Transitivity Coefficient I am much more likely to be friends with the friend of my friend than with a random other person [35, 7.9]. The network property *transitivity* is used to quantify this concept, and also in our network we want to investigate in how many cases it follows from the connection of node i and j and the simultaneous connection of nodes j and k that nodes i and k are connected as well.

The *transitivity coefficient* C is defined as [35]

$$C = \frac{\text{number of closed paths of length two}}{\text{number of paths of length two}}\,, \tag{2.2}$$

where a "path" is a sequence of distinct links that connects distinct nodes and a "closed path" is such a sequence with the additional condition that the start and end point are also connected - sometimes also called a "loop" in network jargon [35]. The length of the path is the number of links that constitute the path. There are a few other closely related definitions of the concept of transitivity coefficient that are also known under the name of "clustering coefficient". A complete graph has perfect transitivity and hence $C = 1$, while some network topologies, such as trees or square lattices have no closed triangles and therefore exhibit a coefficient of $C = 0$. In [35] Newman also gives typical values for the transitivity of different network types, with a social networks ranging from $C = 0.1$ to $C = 0.5$ and technological or biological networks typically having lower values.

The transitivity coefficients of spectroscopic networks are summarised in Table 2.1. A notable fact is the very high transitivity of the network $\mathrm{H_{JB}}$ which can be explained by the fact that the authors only considered a particular set of energy levels, but rigorously studied the transitions between these up to the order of electric and magnetic octupole transitions. The selection of states was shown in Figure 2.1. The transitivity of $\mathrm{H_{NIST}}$ has a value of $C = 0.24$ that is around ten times higher than the transitivity coefficients of helium, carbon and iron. By comparing with the meta data of the networks, we notice that the transitivity seems to be anti-correlated to the share of electric dipole transitions L_{E1}/L in the network. A network consisting of only electric dipole lines would be bipartite due to *Laporte's* selection rule and the transitivity of bipartite graphs is trivially zero. Hence this measure again seems to be connected to the origin of the underlying data of the network and not to the inherent properties of an atomic system.

The Degree Distribution A fundamental property of a network and a defining characteristic of the network structure [35] that cannot be cast into a single number - and hence is not found in Table 2.1 - is its *degree distribution* p_k. Simply stated, the *degree* k of a node is the number of links that it has to other nodes and the degree distribution is the distribution of the probability that a random node of the

network has degree k [2].

In the following we will review if the degree distribution of spectroscopic networks relates them to a general class of network models, *e.g.* rendering them similar to social graphs or random graphs. In particular, we will study if the spectroscopic networks share any resemblance to graphs of the *Erdős-Rényi* model or *scale-free* graphs which are the most common network models.

Erdős-Rényi Model One of the earliest network models is the *Erdős-Rényi* (ER) model, which generates random graphs by connecting an existing set of nodes randomly: each possible link will be included in the graph with an equal probability of p [11]. A common criticism of the *ER* model is that it produces a unrealistic degree distribution which is not found in real networks [2, 3.10]. The term "real network" in this context refers to networks that were constructed on the basis of actual data that is cast into the language of nodes and edges - in contrast to networks that are constructed on the basis of mathematical prescriptions. The degree distribution of an ER network follows the Poisson distribution [2, 3.4].

Barabási-Albert Model A widely spread, yet strongly debated, result of network science is the notion that most real-world networks are *scale-free*, *i.e.* their degree distribution p_k can be described by a power law as a function of the degree k [35, 8.4],

$$p_k = ck^{-\alpha}\,, \tag{2.3}$$

where c is normalisation constant and α is the constant known as the exponent of the power law.

A reason for the wide application of this concept might be the fact that it gives a mathematical reasoning for the existence of "hubs", *i.e.* high-degree nodes, in networks. The abundance of hubs is central in understanding many effects in networks, *e.g.* how susceptible a network is to the spread of an epidemic [32]. In [3] Barabási and Albert trace back the existence of scale-free networks to two mechanisms: growth and preferential attachment. They state that in contrast to the assumptions of the random network model, real networks do not have a fixed number of nodes, but continually grow by the addition of new nodes and that this

addition is such that new nodes preferentially attach to nodes that are already well connected. They proposed the *Barabási-Albert* model which incorporates these assumptions and show how this model leads to the creation of hubs and a scale-free degree distribution.

Regarding the quantum mechanical model of atoms we know that those assumptions do not apply to these systems. However, one could imagine the measuring of a new energy level (transition) and the addition of this information to the existing data as adding a node (link) to the network. Hence a measuring process could be modelled as growth of the network. From this point of view it would make sense that new transitions are more likely to be measured for well known energy levels, similar to the prediction of the preferential attachment mechanism.

Barabási and other members of the network science community studied many more degree distributions of real-world networks to reason that their distributions are power laws and to promote the "scale-free property" as a central characteristic of real-world networks. An often cited example is the study of the internet topology [12]. Because of variety in the types of the networks considered - ranging from social networks, communications networks to biological networks - Barabási calls the scale-free property a universal network characteristic [2, 4.5] and also argues that all scale-free networks are based on preferential attachment [2, 5.9]. Yet there is an ongoing dispute about this universality. It is common empirical knowledge in network science that degree distributions rarely follow a true power law as in equation (2.3) over their entire range: in small k regimes the distribution is often not monotonic due to statistical fluctuations and sometimes even the tails do not follow the power law, *e.g.* when there is a cut-off that limits the maximum value of the degree [6][35, 8.4]. A recent study of almost 1000 networks from different scientific areas disputes the universality claim made by Barabási [6]. A more exhaustive and general discussion of the scale-free property of networks and the mathematical models behind it can be found in Chapter 8 of [35], Chapters 4 and 5 of [2], [6] and the references therein.

In the wake of this discussion we want study if the spectroscopic networks, which are considered to be real-world networks, exhibit the scale-free property. For this it should be mentioned that the study of molecular spectroscopic networks by Császár

and Furtenbacher in [8] and [15] suggested either power-law or log-normal behaviour of the degree distribution, but in any case an abundance of hubs in such networks.

Spectroscopic Networks After having introduced two network models and the most common criticism they are exposed to, we will check if this criticism is justified in the context of spectroscopic networks. We will not go through a very sophisticated fit routine as done in [6], but only check qualitatively, if the spectroscopic networks could be attributed to the class of scale-free networks. This means that we will not use a fit function that incorporates any expected deviations of the degree distribution from the true power law, but only use the power law as stated in equation (2.3). The only extension to this is that for the carbon and the iron network we introduce a lower cut-off k_{min} to exclude the plateaus at low degrees. The data is shown in a log-log plot to portray the power law as a straight line. We also fitted a Poisson distribution to the data to compare it to the predictions made by the *ER* model. The results for each spectroscopic network are shown in Figure 2.2.

Looking at Figure 2.2 we can immediately dismiss the random graph model as a suitable model for this data. The only network for which the Poisson distribution remotely resembles the data is the network H_{NIST}, and even here the peak of the fitted curve is in a local minimum of the data.

With respect to the scale-free property it can be said that no degree distribution considered here follows a power law over its entire range. The distributions of the Carbon and Iron networks seem to follow a power law when low-degree cut-offs $k^C_{min} = 5$ and $k^{Fe}_{min} = 12$ are introduced. However, the degree distributions also show deviations from the power law in the high-degree regime. By this simple analysis it can not be ruled out that these two networks are scale-free and a more rigid analysis following [2] or [6] would have to be carried out for clarification. We also have to consider the fact that usually the exponent α of the power law is said to be in the interval $[2, 3]$ for strict requirements and at least $\alpha > 1$ for weaker requirements of the scale-free property [6]. However, for H_{JB} and the helium network, the exponent α of the fitted power law curve is negative, leading to a positive slope, which is in contradiction to the sparseness of hubs in the *Barabási-Albert* model. The degree distribution of H_{NIST} does not seem to be described well by the power

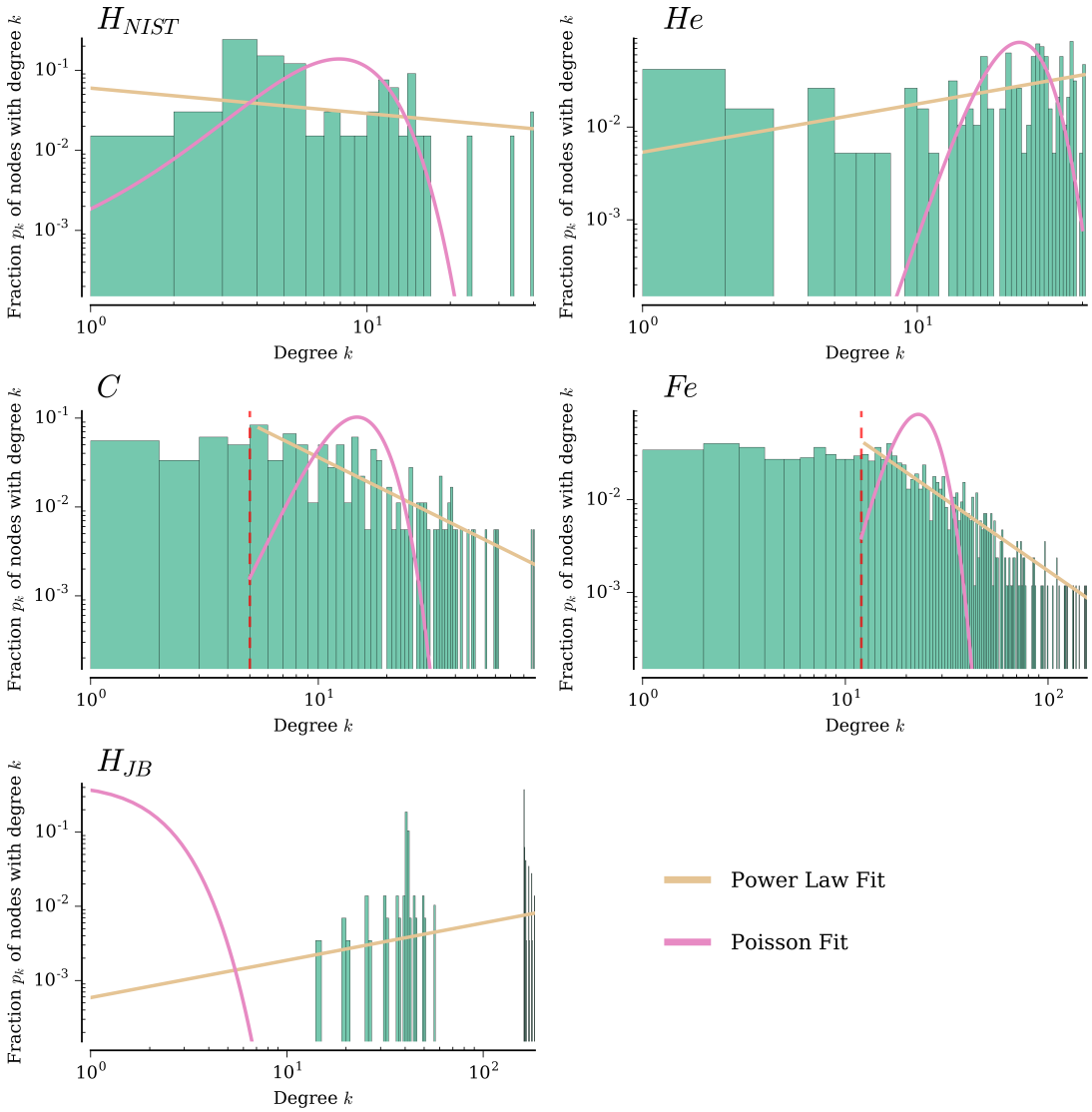

Figure 2.2: Degree distributions of the five example networks. The distributions are plotted on logarithmic scales. For each distribution a power law based on equation 2.3 and a *Poisson* distribution is fitted to the data. The dashed red lines in the distributions of the carbon an iron networks show the respective values of $k_{\min}$, the minimal degree considered for the fit.

law, as the distribution exhibits two peaks that are not captured by this fit function.

Given our earlier observations of differences between the experimentally and theoretically obtained networks and the possibility to model the measuring process of the spectroscopic data as the growth of a network, the bias introduced by the selective inclusion of transitions in the data might be the reason that the degree

distribution of spectroscopic networks resembles the one of scale-free networks. H_{JB} and the helium network do not show this behaviour, yet they are primarily based on theoretical data: the biggest source of the data in the NIST data base for helium are theoretical calculations in [9] where only electric dipole transitions were considered. The H_{NIST} network is the smallest network and the deviations from the power law behaviour could be attributed to statistical fluctuations.

Thus, while it can be ruled out with certainty, that the spectroscopic networks are based on the *Erdős-Rényi* model, the experimental evidence is less clear with respect to the *Barabási-Albert* model. However, even if these networks had the "scale-free property", we argue that this would be due to the particular way their data was obtained and not due to inherent properties of the atomic systems - which we are ultimately interested in. This argument is not only supported by the empirical evidence provided here, but also by our quantum mechanical understanding of atoms. This suggests that for spectroscopic network, the degree distribution is not a characteristic function that leads to insights about their underlying physical processes.

Degree Assortativity Coefficient To uncover the building principles of the network we want to study how one nodes property influences another node, *i.e.* how they might be correlated. Networks can share the same degree distribution but have quite different structure, *e.g.* depending on how likely hubs are to connect with other hubs [35]. If we find such a non-random trend of connections in the networks, we might be able connect this information to a physical origin.

In this work we are not using any data other than the network's composition, so we do not have any node properties like quantum numbers or the energy of the state that would be useful in the context of atomic systems. Yet luckily a scalar quantity that comes in a package deal with the structure of the network - and that we thus can always rely on as a node property - is the node degree: the number of links a node is connected to. The availability and genericness is one reason why we will discuss the assortativity with respect to the degrees of the nodes. Another reason why assortativity by degree is particularly interesting is that the degree itself is a structural property. If the degree dictates the position of the edges in the networks, this leads to interesting features [35, 7.13.3].

One possibility to describe the degree assortativity of nodes is to measure the covariance of the value pair (k_i, k_j), where k_i is the of node i, for all edges (i, j) in the network [35, 7.13.2]. The covariance will be positive if the degrees of the both nodes of a link tend to be both large or both small and it will be negative when they tend to oppositional values [35, 7.13.2]. Normalising by the maximum value of the covariance leads to the so-called *degree assortativity coefficient* [35][33]

$$r = \frac{\sum_{ij}(A_{ij} - k_i k_j/2L)k_i k_j}{\sum_{ij}(k_i\delta_{ij} - k_i k_j/2L)k_i k_j}, \tag{2.4}$$

where L is again the number of links in the network and A_{ij} are entries of the adjacency matrix. The adjacency matrix $A \in \{0, 1\}^{N\times N}$ is a matrix representation of the network with $A_{ij} = 1$ if a link between node i and node j exists and $A_{ij} = 0$ otherwise. It is a convenient way to store and handle networks, especially on computers [35]. We use the symbol r for the degree assortativity coefficient to not confuse it with the adjacency matrix A and because of its relation to the Pearson correlation coefficient [35]. The degree assortativity coefficient is only dependent on the network structure and can be calculated solely from the knowledge of the adjacency matrix A. In this form it varies in value between a maximum of 1 for a perfectly assortative network and a minimum of -1 for a perfectly disassortative one; $r = 0$ indicates that the node degrees are uncorrelated [35]. Calculating the degree assortativity coefficient for the spectroscopic networks leads to the values shown in Table 2.1.

Newman also states that there is a natural bias for simple, *i.e.* undirected, graphs to show disassortative behaviour which is typically overcome by social but not by technological or biological networks [35, 7.13.3]. The spectroscopic networks also show this disassortative behaviour, with the H_{NIST}, carbon and iron network all having negative assortativity coefficients. Yet the strong assortative behaviour of the helium graph prohibits a definite conclusion about the general behaviour of spectroscopic networks with regards to assortativity. A possible reason for the disparity of the values for the different networks could be that focus of the publications, which are the sources for the data, introduced a bias which is masking the inherent behaviour of the networks. We previously stated that the biggest source of the data in the NIST data base for helium are theoretical calculations. The focus on some energy states and their transitions, while just sparsely studying other states introduces an

assortative *core-periphery* structure. This would explain helium's assortativity coefficient of 0.43, which is the only positive value for networks based on NIST data. The same applies to the H_{JB} network due to the biased selection of energy states, as shown in Figure 2.1.

2.2.2 Comparison of Hydrogen Networks

In this section we will make a side-by-side comparison of the two networks that we have got for the same physical system: the hydrogen atom. As previously stated, one of these networks was generated by data that was obtained by experimental observation and is collected and made available in the NIST database [25][26]. This network is also referred to as H_{NIST} and has 66 nodes and 287 edges. The second network is based on the work of Jitrik and Bunge in [22]. We have already depicted in Figure 2.1 which states the authors considered therein. They derived wavelength and transition probability for all atomic transitions up to the order of electric and magnetic octupole between the considered states. Based on this data we constructed a network consisting of 289 nodes and 15 793 edges. Since both networks describe the same physical system, the graphs should be ideally the same. Yet, by way of construction, the theoretical network H_{JB} is complete up to the order of electric and magnetic octupole transitions (for the states considered), whereas the H_{NIST} network lacks many of these such transitions. We want to study how the experimentally obtained data differs from this theoretical network.

A reason for the omission of transitions in H_{NIST} could be the experimentally motivated focus on selective lines that are only measured when needed for a scientific investigation (nobody will measure thousands of transitions without a good reason). This would have created a bias in the experimental data, impacting the network structure, changing the outcomes of the network analysis and its interpretation.

Obviously it is not sufficient to solely compare the two hydrogen networks and assume that the other atoms have the same biases as the experimental data of hydrogen. Different elements are put to completely different scientific and practical use, but it is only for hydrogen that we are given such an extensive network which is based on theoretical calculations. The majority of the transitions in the helium network are also theoretically obtained, but these are all of the electric dipole type

(E1). This comparison is less a statistical analysis of the structural differences, but rather an exploration of the network features aiming for an intuition that will help in the interpretation of later results.

Basic Properties The most prominent difference between these networks is their respective size. The theoretical network H_{JB} has more than 4 times as many nodes and more than 50 times as many links as the experimental network H_{NIST}. As the number of possible links L grows proportional to the number of nodes N squared, this leaves H_{JB} with a higher density $D_{H_{JB}} = 0.35$ than H_{NIST}, for which $D_{H_{NIST}} = 0.11$. The differences in density and clustering can be attributed to the way of construction of H_{JB}, as done in Section 2.2.1.

The difference in the degree assortativity coefficient of the networks has also been mentioned in Section 2.2.1: the nodes in H_{JB} tend to connect to nodes with similar degree to their own, while, in contrast, the nodes in H_{NIST} prefer to connect to nodes with oppositional degree values (meaning that high-degree nodes tend to connect to low-degree nodes instead of other high-degree nodes). A possible explanation for the disassortative mixing in H_{NIST} is the inclusion of spectral lines from the lowest energy states to many highly excited states. Although the low energy states are hubs and their connection to other states has been well investigated, the highly excited states are only considered in these particular transitions and their connection to the remaining state was apparently of no greater importance, rendering them low degree nodes in this network. Furthermore, some of these hubs cannot be connected because of constrains due to the selection rules, which further decreases the assortativity.

The differences between the networks have so far also been impacted by the fact that different energy states were considered in the respective networks. To put the two networks on equal footing when discussing their structural differences one should actually consider only energy states that occur in both networks. Table 2.2 lists thus the basic properties of H_{NIST} (as in Table 2.1) and of the subgraph of H_{JB} that only includes nodes also occurring in H_{NIST}. One can see in this table how the properties of this subgraph resemble the properties of a complete graph, with the values of density, assortativity and clustering coefficient all being close to the

respective values for a complete graph. Especially the degree assortativity coefficient has changed with respect to the full H_{JB} network.

Table 2.2: The network properties of the subgraph of H_{JB} that only includes energy states in the data set for H_{NIST} in comparison with those of H_{NIST} and a fully connected network (a complete graph) with the same number of nodes. For a description of the properties see caption of Table 2.1.

Network	N	L	D	r	C
H_{NIST}	66	278	0.18	−0.31	0.49
H_{JB} (subgraph)	66	1 905	0.89	−0.008	0.94
Complete Graph	66	2 145	1	0	1

Transition Probabilites The Schrödinger equation allows us to calculate the transition probability between two quantum states. If the matrix element in the calculation of such an electron transition vanishes in the dipole approximation, the transition is said to be forbidden [5, 4.3]. The *selection rules* state the cases for which this happens in terms of the quantum numbers of the respective quantum states [5]. The above notion of "forbidden" is better described as "electric-dipole-forbidden", as the consideration of further terms in the expansion of the matrix element gives rise to higher-order transitions whose transitions rate is non-vanishing, such as magnetic dipole transitions, electric quadrupole transitions *etc.* [5]. These higher-order transitions are considered in H_{JB}, yet their transition rates are disregarded in the simple graph model, as in this model a link either exists or does not.

We argue that for describing an atomic system, it is desirable to include also the higher-order transitions in the data, as in principle done in the network H_{JB}. However, without considering "link weights" of any kind, when including more data the properties of H_{JB} will further approach the properties of a complete graph and the network loses its ability to convey the structure of the atomic system: if no differentiation between the links is made, a complete graph does not store any structural information besides its size. Hence, to capture the structural behaviour of the atomic systems, when using the particular representation of states as nodes and transitions as links that we do, we ideally should have a complete graph and the structural properties should be encoded in the link weights, given by the transitions rates.

In general this is not possible for us, since we do not have enough data about these higher-order transitions for most of our networks. Also, if we were to include them, the transition rates greatly vary in value over many orders of magnitude and this is an obstacle for many numerical algorithms that are routinely used in network science (although a lower cut-off could be part of the solution for this problem; G. Czászár and Furtenbacher used this practice in [16]). Thus, we are restricted case of simple graphs in which we have not included many lines that are not electric-dipole transitions (as indicated by L_{E1}/L in Table 2.1) and when they are included, we disregard differences and treat all links equally. However, using these unweighted, experimentally obtained network does not mean that the data is unusable or unsuited to describe the system. Because of the experimental focus on lines of high intensity or special interest, the selection of data can be seen as generating a "disturbed threshold".

To study the extent and nature of this disturbed threshold, in addition to the previous measures, we will compare the data of $\mathrm{H}_{\mathrm{NIST}}$ and H_{JB} with respect to the included link weights. Explicitly this means that we calculate the degree, the sum of weights and the maximal weight corresponding to each node in $\mathrm{H}_{\mathrm{NIST}}$ and H_{JB} and visualise the ratio of the values in the respective networks as a box plot. Therefore we also extend our network model for remainder of this section and consider the transition rates as link weights. Figure 2.3 shows the above mentioned properties on the left side of each plot. We see in that there are almost 4 times as many links with known weights in H_{JB} than in the $\mathrm{H}_{\mathrm{NIST}}$ network. Yet this does not change the sum of weights in the network by much, indicating that mostly higher-order transitions are missing from the $\mathrm{H}_{\mathrm{NIST}}$ data, as their weights only make small contributions to this total sum. We see a similar picture when only the maximal weight of each node is considered, although the variance is higher in this case. This means that some transition lines with the highest transition rates are missing from the experimental data set.

One naïve explanation for the fact that lines with high transition rate seem to be missing could be that the experimental data primarily considers the visible part of the total wavelength spectrum and therefore some transitions with substantial contributions were neglected. Therefore, in Figure 2.3 we also study how the network changes when we only consider the transitions that are within a particular wave-

length interval, this is shown on the right side of each plot. This interval is chosen such that it roughly resembles the visible spectrum (it is chosen a bit wider to avoid disconnecting too many nodes by excluding their only link), to check if an experimental focus on this interval is a possible explanation for the omission of lines. As a matter of fact, after applying this filter, the networks are seemingly identical with respect to the sum of weights for each node and thus also the respective maximal weights for each node. However the number of links of the nodes in H_{NIST} is still on average about half the value of the nodes in H_{JB}.

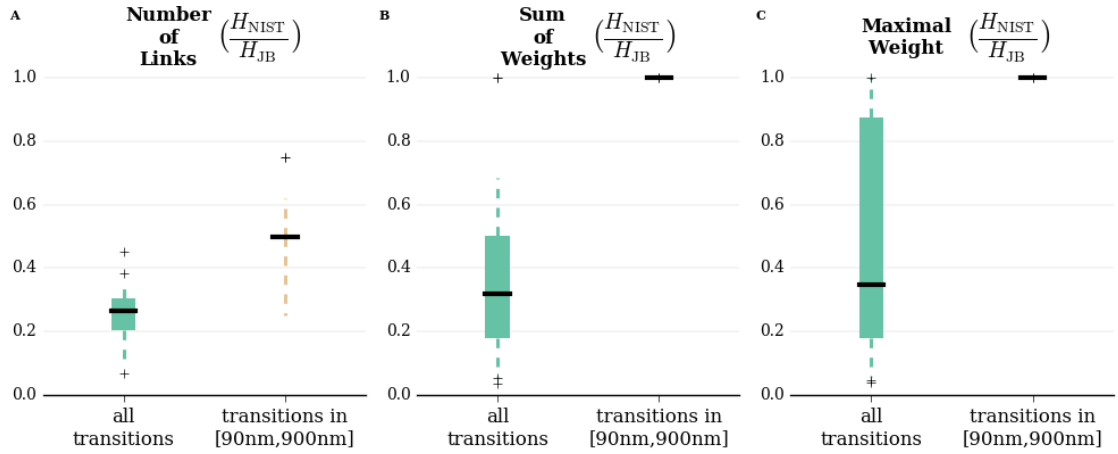

Figure 2.3: Comparison of weight-related node properties in H_{NIST} and H_{JB}. The ratio of each value in H_{NIST} over the respective value in H_{JB} is calculated for each node and the values are subsequently arranged in box plots. The solid box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers (dashed) extend to the 5th and 95th percentile, respectively. The right box plot in each figure considers only transitions whose wavelength is in the range of 90 nm to 900 nm. Transition rates were not given for all spectral lines in the data and such unweighted links were excluded.

2.3 Conclusion

Network science has historically been driven by the study of social networks, and many network properties have therefore been developed in that context, but subsequently also proved to be useful in other network types. However, when we apply these general concepts to spectroscopic networks we notice that they have little explanatory power. It is obvious that for each network type and each research goal

different questions have to be asked, and this type of network seems to be special in the sense that the simple network metrics seem not to be adequate and effective to learn about the underlying physical system.

The network density is affected by the experimental interest and dependent on whether higher-order transition types were included in the data. Neither clustering coefficient nor degree assortativity coefficient seem to lie in any particular range that could in retrospect be called a characteristic value range for spectroscopic networks. The degree distribution has in some cases a form similar to a power law. However the degree distributions of the studied networks differ such that it cannot be said that spectroscopic networks can be attributed to a particular class of networks nor that they share a common kind of degree distribution. A fundamental difference to other network types could be the reason for this behaviour: the nodes are not "agents" like in social networks. There is no component in the structure of the network that is of human origin. Such a human factor only comes into play when we consider the measuring process of the data as a dynamical process of the network. Then we can model this dynamics as growth of the network. Such a point of view could explain the observed differences for the degree distributions. However, if we choose such a model, we would then study not the atomic system itself, as we set out to do, but merely the experimental evolution of the data.

We also discussed an idealised case of representing the structural properties of this particular network type. In this ideal we would encode the transition probabilities in the link weights of a fully connected network. This is not possible to realise because of the incomplete data that we have for most atomic systems, as it is usually experimentally obtained. We established how the experimental data taking leads to a "disturbed" or "noisy threshold", so that the data typically features only the transitions with the highest transition probabilities (*i.e.* mostly transitions of the electric dipole type). Instead of having complete knowledge about the network structure, we can see the simple graph representation that we are using as a first approximation to the system and propose to use methods from network science to extend our incomplete knowledge. Thus we will show in the subsequent chapters that this representation is sufficient to gain insights about the quantum mechanical system and even how to make informed predictions about transitions that are missing from the data.

In the light of this discussion, we also compared the two networks and data sets that we possess for the hydrogen atom. We established the main differences between these data sets and how they impact the network structure in this representation.

It should also be mentioned that by having such an extensive and theoretically obtained first-principles network as the H_{JB} network is, we have a microscopic model of these systems that enables us to validate some of our findings. It is uncommon in network science to have such a detailed knowledge about the underlying system that a network is describing and we will often base our reasoning on this knowledge.

3 Community Detection

In the preceding chapter we have looked at some fundamental network properties of the spectroscopic networks, and tried to connect them to the physical properties of the atomic systems. However, it seems that these properties were heavily influenced by how the experimental data was obtained and that the complexity of the systems could not be grasped by the simple approaches of the network properties.

Another tool to discover and understand the structure of a network are community detection algorithms. Community detection is the search for groups of nodes that are naturally occurring in the network [35]. It is an example of an unsupervised machine learning task as its goal is to classify unlabelled data by using structure in the data that we are initially unaware of. It is then the researchers task to understand and determine the responsible features or processes for the observed partitioning. This is the promise of community detection: the potential to gain valuable insight into the central aspects that govern the function and evolution of the network by seeing its building blocks. Since community detection is an unsupervised learning task, there is no precise definition of what constitutes a community [42]. Different community detection algorithms will therefore be reflecting different notions of what a community is.

We already established in the last chapter that we are not considering any kind node properties in the spectroscopic networks, because we are focussing on simple graphs in this work. Therefore we will discuss approaches for community detection based exclusively on the structural data of the network.

In this chapter we will give a brief introduction to community detection algorithms to be able to discuss the results in the light of its strengths and shortcomings and seek to answer the following question: Are there meaningful node communities in spectroscopic networks? If so, how does this pattern connect to the properties of

atomic systems?

A more detailed treatment of this part of the research project can be found in the final thesis of my colleague D. Wellnitz.

3.1 Methods for Community Detection

One of the earliest and most widespread methods for community detection is the modularity optimization method [36] that finds only partitions of nodes for which the fraction of internal links inside each community is larger than expected in a random graph [38]. The mathematics behind this method is similar to the degree assortativity we discussed in Section 2.2.1. However it suffers from serious drawbacks, such as a tendency to balance the size of communities, its resolution limit regarding the size of communities, the degeneracy of its results for large networks and its incapability to provide statistical evidence for the deviation from the null model, *i.e.* it cannot differentiate between actual structure and statistical fluctuations [38].

To overcome the limitations of the modularity method, a more statistically rigorous approach was developed. A current state-of-the-art approach is the nested stochastic block model developed by Peixoto [38]. This approach generates a grouping of nodes based on non-parametric bayesian inference and estimates the likelihood of such a grouping based on an information entropy measure. The advantages of this approach are the following [38]: it generalises earlier approaches and can describe different types of structure - not only assortative structure, which assumes that communities are mostly connected to themselves, but also disassortative and hierarchical structure. It has a lower resolution limit than the modularity based method. As it is using concepts of information entropy and is non-parametric, it is looking for the simplest way to describe the data and is thus not prone to over-fitting data. Peixoto also claims that it scales well for very large networks [38]. For the following results, this method has been adopted for the investigation of spectroscopic networks.

A detailed description of stochastic block models can be found in [39] and with a focus on the nested stochastic block model in [38]. For a more bird's-eye-view discussion of community detection see [42] and [14].

3.2 Communities in Spectroscopic Networks

As an example for the result of the algorithm [38], Figure 3.1 shows the nodes of the helium network spatially arranged according to the 20 communities that the method found.

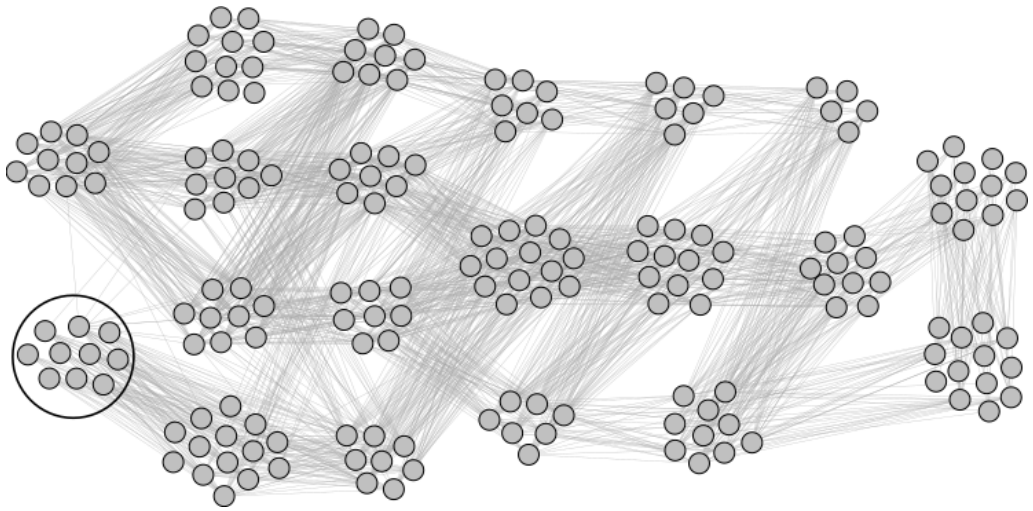

Figure 3.1: Depiction of the spectroscopic network of helium. Nodes represent the energy states of the atom. Two nodes are connected by an edge if there exists a transition between the two corresponding states that the nodes represent. The nodes are spatially grouped by the results of the nSBM community detection algorithm. A meta data analysis reveals that all states of the circled community have the same values of the L, J and S quantum numbers. This is also the case for another 11 communities.

To find the patterns or features by which the nodes have been grouped, we compared the network's meta data with the community structure. We notice that all energy states that the nodes in the circled community represent share the same total angular momentum quantum number j, azimuthal quantum number l and spin quantum number s. We further see that this is also the case for 11 of the 19 remaining communities (or 13/19 if we disregard doubly excited states, explained in detail in the next section).

A one-electron atomic state is defined to sufficient accuracy by the quantum numbers n, l, j, and m_j [31]. n is the principal quantum number and describes the shell of an electron; the azimuthal or angular quantum number l describes which *subshell* (also "*orbital*") an electron "sits in"; the magnetic quantum number m_j describes the projection of the total angular momentum along a specified z-axis

[5][31]. Taking into account relativistic corrections, we need to include the quantum number j which is connected to the total angular momentum of the electron and in one-electron systems $j = l \pm 1/2$. For systems with several electrons, there is also a multiplicity of the states with respect to the spin, because the spins $\mathbf{s_i}$ of the single electrons are coupled to give a total spin angular momentum $\mathbf{S} = \sum_i \mathbf{s_i}$. Thus we have to explicitly include the spin quantum number S [5]. To highlight that the quantum numbers are describing the collective of electrons, we use capital letters from now on. Inconsistently we do this as well for states of hydrogen, although this system only has one electron.

In the following we seek to quantify this correlation between the group membership of nodes and the quantum numbers of the states.
If the network were completely split according to the L, J and S quantum numbers, each group would correspond to a collective of *fine structure components* characterised by this set of quantum numbers [5]. Note, however, that they would be the fine structure components of different shells as we disregard n. Since we do not know, to what extent the communities resemble the fine structure of the atoms, we want to quantify the performance with respect to different combinations of the quantum number set. Thus we define different "ground truths" for the nodes' labels: a ground truth given by only the quantum number L, one given by only J, one given by the combination of L and J, one given by the combination of L and S and one by the combination of L, J and S. For the hydrogen networks we will not use the last two combinations, as it is a one-electron system and thus S is the same for all states. As stated earlier, we disregard magnetic quantum numbers as we do not consider the presence of external magnetic or electric fields. Also we disregard the principal quantum number n in our ground truths: as the selection rules (for electric dipole transitions) are not dependent on n, the network structure is not sensitive to this quantum number and the communities are hence degenerate with respect to it. If we were to implement the explicit energies in the model, *e.g.* by using energies as link weights, we could break this degeneracy with respect to n.

A common metric to quantify the accordance of two sets of clusterings is the *Rand index*, that measures the number of agreements of the two sets compared against the number of disagreements [41]. However, we are not using the Rand index in its original form, but the *adjusted Rand index* (ARI), which has an expected value of

zero for two independent clusterings, and a maximal value of 1 for identical clusterings [21].

In Table 3.1 we have listed the accordance of the output with the distinct ground truths measured by the adjusted Rand index. For this, the ground truth was compared with the best partitioning we could find according to the likelihood from the nSBM method. The results in this table show that the correlation between the studied quantum number sets and the communities found is the highest in the helium network, but is also present to a lower extent in the remaining networks. In the H_{NIST} network the correspondence of the communities with the ground truth given by L has an ARI of 0.165, while we find a value of 0.027 for the J and 0.069 for the LJ ground truth. We discussed in Chapter 2 how the experimental selection of data influences the network structure, as this selection is not strictly guided by a physical parameter like the transition probability. It seems that because of the small network size and this disturbed threshold, the algorithm was not able to properly learn the atomic fine structure from the structural data. It is not surprising that the ARI value is higher for the L ground truth than for the J ground truth, as the selection rules for L are more strict (fewer values are allowed for ΔL than for ΔJ in dipole transitions).

The values for the helium network are as follows: the ARI value for the agreement of the L ground truth and the community structure of helium is 0.409. In contrast to the hydrogen networks the value for the J ground truth is approximately equal at 0.419. This indicates that this time both selection rules could be found. The LS ground truth has an even higher agreement, indicating, that for this network also the separation into singlet and triplet states was found to some extent. The ARI value for the LJ and the LJS ground truths are both 0.791. The reason for this high agreement might be twofold: firstly, this light atom is well described by these quantum numbers. Secondly, because the data is mostly from a rigorous theoretical calculation of electric dipole lines this network has less emphasis on the higher-order transitions than the other experimental networks that are also based on NIST data. Hence the structure is mostly driven by the selection rule set of the electric dipole transitions and there are only few disturbances by the selection rules of the other transition types or from the omission of transitions. That the ARI values of the LJ and LJS ground truths are equal (at this precision) suggests that the separation

into singlet and triplet states was not found consistently throughout the network. Otherwise the value for the agreement with the LJS ground truth would be significantly higher than for the LJ ground truth.

In the H_{JB} the ARI values are higher than for the H_{NIST} network, with 0.289 for the L ground truth and 0.316 for both the J ground truth and 0.178 for the LJ ground truth. Although this network is solely based on theoretical calculations, these were carried out for all transition types up to magnetic and electric octupole transitions. Hence, in contrast to the helium network these transitions are represented more evenly, leaving the algorithm little room to learn discriminating features. The higher value for J might be explained by the fact that the selection rules of the higher-order transition types include the rules of the dipole transitions with respect to J, *i.e.* the octupole transitions do not violate the dipole selection rule concerning J.

The communities in the carbon network have a significantly lower agreement with the ground truths compared to the helium network. The highest ARI value of 0.214 is found for the LJS ground truth. This decrease is most likely attributed to the increase of the systems complexity. As carbon has a higher number of electrons there are more correlation effects, making the quantum numbers less appropriate to describe the system in its entirety. Consequently the ARI values decline more for the iron system: the highest agreement is for the LJ ground truth with an ARI of 0.150.

The results also show that for the lighter atoms it is easier to find the L quantum number, whereas the network structure of the heavier atoms seems to be more reliant on J. This matches the interpretation that coupling effects are increasingly important in those atoms. While L loses in importance as the parity of a state is no longer defined by the value of L, the quantum number corresponding to the total orbital angular momentum $\mathbf{L}$, but by the sum of the l values of the single electrons. The J quantum number stays appropriate for heavier atoms, but it is obviously not sufficient on its own to describe the states and therefore even the values for the J ground truth are decreasing substantially for these systems.

Table 3.1: Agreement of the network clusterings by the nSBM algorithm and the clusterings according to different quantum numbers (ground truths) measured in terms of the *adjusted Rand index* for each example network. The adjusted Rand index has an expected value of zero for two independent clusterings, and a maximal value of 1 for identical clusterings. The ground truths are the quantum numbers of the energy states orbital angular momentum (L), total angular momentum (J), spin (S) or combinations thereof. There are no errors given to the values, as solely the partitioning with the highest likelihood - as proposed by the nSBM algorithm - was used. The highest value for each network is in boldface. See Table 2.1 for an overview of the example networks.

Network	Ground Truth Set	Adjusted Rand Index
H_{NIST}	L	**0.165**
	J	0.027
	L, J	0.069
H_{JB}	L	0.289
	J	**0.316**
	L, J	0.178
He	L	0.409
	J	0.419
	L, J	**0.791**
	L, S	0.487
	L, J, S	**0.791**
C	L	0.039
	J	0.149
	L, J	0.152
	L, S	0.119
	L, J, S	**0.214**
Fe	L	0.044
	J	0.123
	L, J	**0.150**
	L, S	0.075
	L, J, S	0.115

Community Structure of Helium

In the last section we established that for the helium network the communities that were found by the nSBM method have an accordance of 0.791 with the communities according to the three quantum numbers L, J and S when measured by the adjusted Rand index. We have given some intuitive explanation into why this result excels the one in the other example networks. Yet from the ARI value we also notice that there is room for improvement. In the following we will discuss the results regarding the community structure of the spectroscopic network of the helium atom in more detail to study the errors made by the nSBM method. If there are patterns in the occurrence of errors, we can deduce how to avoid them in the future.

In Figure 3.2 we see again the helium network with its nodes spatially grouped by their corresponding group membership as given by the nSBM method. Additionally the nodes are labelled by colour and sign according to their classification by the *Russell-Saunders* notation ${}^{2S+1}L_J$, an abbreviated description of the states' angular momentum quantum numbers L, J and S. The black nodes in the graph are doubly excited states of helium with energies above the ionisation threshold. These states can either decay to bound states of helium via a radiative transition that does not violate the selection rules or decay into a free electron and an He^+ ion via a radiationless Auger transition [5]. Since the latter is not captured in this model we disregard these states in the further discussion.

A first glance at the network graph reveals that there are 30 distinct fine structure term symbols ${}^{2S+1}L_J$ (*i.e.* combinations of L, J and S) by which the nodes are labelled. In 14 instances, all states with the same fine structure term symbol have been grouped together without omission of an alike state or intrusion of a state with a different label. Looking at the communities with mistakes, one notices that there is never separation of states with the same label, but only mixtures of labels; *i.e.* states with the same term symbol are always members of the same community. The mistakes are always such that the communities are not split up enough. The errors therefore result from under-fitting the network data.

It is also easy to notice that the flawed communities are all on the right side of this graph. Since the nodes are drawn with increasing L quantum number from

the left to the right, *i.e.* the errors are all happening for groups with high L. The first mixture of energy levels happens for the $^1F_3$ and $^3F_3$ levels (depicted in yellow) where the algorithm makes no distinction between the singlet and triplet states of helium. At this point the method loses the ability to distinguish states by means of the S quantum number. The graph shows that many intercombination lines from $^1D_2$ states to $^3F_3$ states are in this data set and thus the $^1F_3$ and $^3F_3$ states appear equal from a network point of view. The same mistake happens for the $^1G_4$ and $^3G_4$ states (green) as well as the $^1H_5$ and $^3H_5$ states (turquoise). Further, the states with the $^3G_5$ fine structure term symbol are mixed together in a community with the $^3I_5$ states. This is the first instance (by ascending L) of mixing states of different L subshells. Nonetheless, these states all have even L quantum numbers and are thus harder to distinguish, as their parity is equal. The remaining states of the I shell have all been assigned to one group. The last community consists of the states of all 4 different fine structure term symbols of the K shell mixed with the $^3H_6$ states. This group therefore breaks all symmetries except parity, which is never broken in this grouping scheme.

Summarising the explicit errors, we notice that for terms with higher L quantum numbers, we first lose the separation of levels with different S, followed by L, while the J quantum number is conserved the longest and the parity symmetry is never broken in this partition scheme.

The increasing error rate with increasing L could be explained by the decreasing number of nodes in this regime, as fewer states were studied for the excited terms. It is then easy to follow the argument, that the lack of data in this regime introduces more errors as the algorithms' ability to classify heavily dependent on the amount of usable data. The likely cause for the mixing between singlet and triplet levels (*i.e.* loss of separation with respect to S) is the presence intercombination lines while simultaneously treating all links with equal weights. Considering transition probabilities would lead to a more nuanced network structure, aiding in a the distinction between the levels of orthohelium and parahelium .

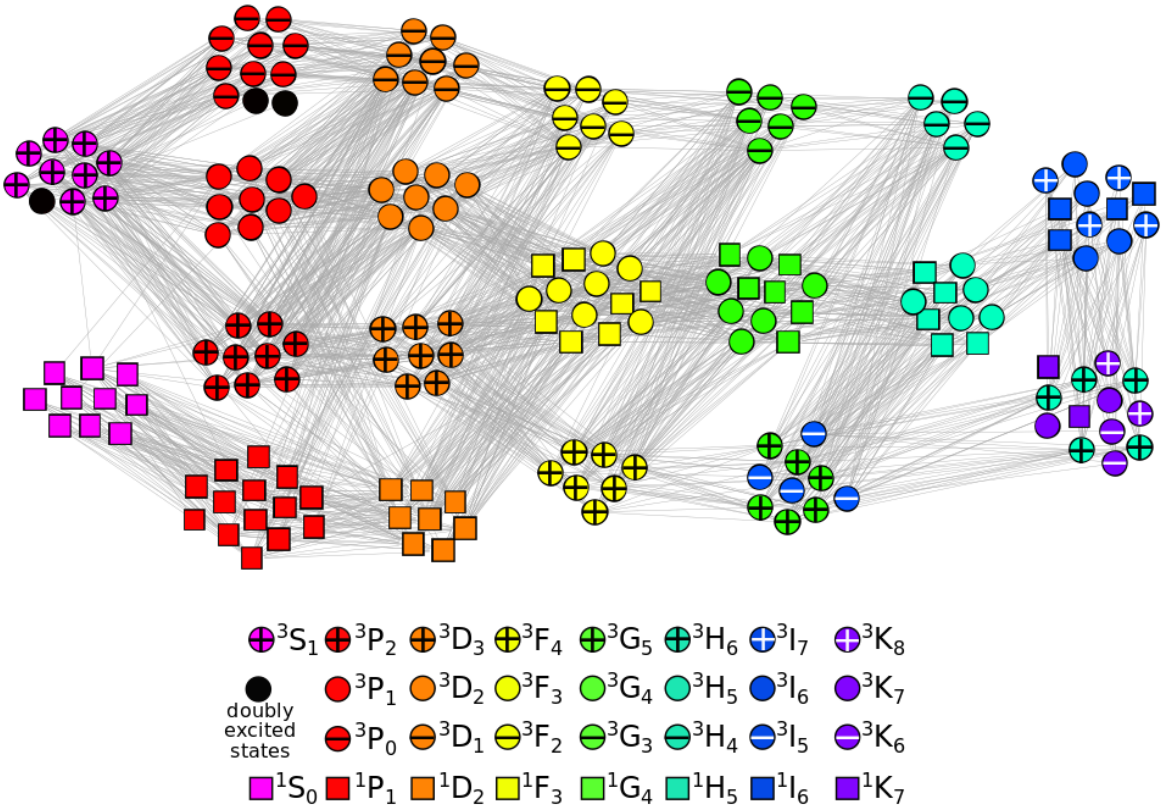

Figure 3.2: Depiction of the spectroscopic network of helium with node labels corresponding to the fine structure term symbols in the *Russell-Saunders* notation $^{2S+1}L_J$ (see figure legend). Nodes represent the energy states of the atom. Two nodes are connected by an edge if there is a transition between the two states. The nodes are spatially grouped by the results of the nSBM community detection algorithm. The algorithm separates nodes into groups with similar quantum numbers. Up to $L = 2$ the communities are separated into states with equal L , J and S, with the only exception being three unstable doubly excited states (black). For $L \geq 3$ the singlet and triplet states are grouped together and for $L \geq 4$ mixing between states with different L occurs. States with different J are mixed for $L \geq 6$. States in the same group never have different parity.

3.3 Conclusion

We saw in this section that the state-of-the-art community detection algorithm nSBM [38] was grouping the nodes of spectroscopic networks according to the quantum numbers of the energy states. This method was able to connect the network structure to physical quantities of the atomic systems - in contrast to the basic network properties of Chapter 2, which were mostly connected to the origin of the data. Thus, in contrast to the previous results, in this chapter we demonstrated the connection of network features with physical quantities. Although no new insights into atomic systems were obtained, the agreement of the results by this community detection method with the quantum mechanical theory of atoms can be seen as a proof of concept for the study of atomic systems by the means of network theoretical methods.

However, one has to note that this interpretation is primarily based on the helium network, as this was the only system for which a thorough analysis was possible. The other networks were too small (H_{NIST}), too biased (H_{JB}) or did not have a sufficiently well known quantum mechanical ground truth (C, Fe). It is also possible, that some results are corrupted by artefacts that derive from the fact that these datasets were measured by different academic groups which might have had different foci in their experiments. For these reasons, the results are not clear without ambiguity and leave room for alternative interpretations for single values of table 3.1. However, the general trends in the values tally with our knowledge about atomic systems.

From the detailed analysis of the differences between the communities found and the ground truths set according the quantum numbers L, J and S, we argue that more data as well as introducing link weights would improve the results. Yet we are so far unaware of a feasible method that includes link weights with a value range of several orders of magnitude in its search for communities.

We ought to emphasise again that other community detection algorithms with a different notion of communities could lead to different results. This would not render the results of this chapter any less significant, but an additional meaningful community structure would rather increase the relevance of spectroscopic networks.

It is also notable that the nested stochastic block model method allowed it to find the disassortative communities of the system, demonstrating the generality of this approach.

Altogether it is astonishing that this approach was able to recover the underlying symmetries of the physical system without ever using quantum mechanical properties, but solely by means of the structural data of a network. The symmetries of the physical system are encoded in the community structure of the spectroscopic network since the selection rules dictate the placement of their links.

4 Link Prediction

So far our goals have been discovering and understanding the structure of the spectroscopic networks and bringing the results into accordance with our knowledge from quantum mechanics. A new set of tools - common methods from network science - enabled us to get insights into the concept of quantum numbers without using the microscopic theory of atoms. The results of the preceding chapter showed that the mapping of the atomic spectra data onto a network leads to results that are indeed backed up by our current understanding of quantum mechanics.

In this chapter we want to go further and use the network approach to tackle tasks that are currently not feasible by means of quantum mechanical calculations. Exact calculations of quantum mechanical systems are only possible for small Hilbert spaces due to the exponential dependence of the number of particles. The current approximative approaches have a variety of limitations (see [23] and references therein), so that at the time of writing no satisfactory treatment of atomic systems of higher atomic number is possible. By using the information that is encoded in the structure of the network we can predict transitions, that are not yet experimentally observed, without using any knowledge of quantum mechanical theories. Such a method can be used as a tool in experimental spectroscopy.

We seek to address the following questions in this chapter:

- Is it possible to predict unobserved atomic transitions by using only the network structure?
- If so, which method is best suited to tackle this problem in spectroscopic networks?
- How can we validate the predictions and quantify the performance when we have no way to check the results by exact theoretical calculations?

These questions will be addressed in the framework of link prediction, a problem that has been widely studied in network science. Thus, in this chapter we will introduce the concept of link prediction, several algorithms that output such predictions and study how they fare in the context of spectroscopic networks.

4.1 The Link Prediction Problem

For the following discussion we will again restrict ourselves to the case of an unweighted and undirected network, *i.e.* a simple graph, $G = G(V, E)$ with nodes $v \in V$ representing the energy states of an atom and the links $e \in E$ representing the atomic transitions between two states. Furthermore let $|V| = N$ and $|E| = L$. Suppose our knowledge of G is complete with respect to the nodes, *i.e.* we know all nodes in V, but incomplete with respect to the transitions occurring, *i.e.* we only know links in a subset E^T of E while we are unaware of the complementary subset E^P and its size. Hence $E = E^T \cup E^P$ and $E^T \cap E^P = \emptyset$. Additionally, let U be the "universal set" containing all possible links between the N distinct nodes. It is easy to see that $E \subset U$ and the cardinality of the universal set is $|U| = \frac{N(N-1)}{2}$ [30]. The task of link prediction can then be formulated as follows: Find the links in the set of unobserved links $U - E^T$ that are members of the set E^P (links that exist, but are not observed due to our incomplete knowledge of G) [30]. Since we do not know the size of E^P we are unable to give an explicit list of members, but can rank all links in $U - E^T$ according to their likelihood of existence. This list of predicted links and their corresponding likelihoods of existence is the primary result of link prediction algorithms. Figure 4.1 illustrates the link prediction problem in a small example network.

In the following the performance of different link predictions algorithms on spectroscopic networks will be examined. For this we need to specify how we validate the predictions: as we want to predict new links, we do not know if they exist or not. Therefore we do not have an absolute ground truth of the network that we can use to validate the results of the algorithms. Furthermore, we need to define an appropriate measure for the performance of link prediction algorithms.

Above we assumed that there are no false positives, *i.e.* no spurious links, in the network. For an approach considering this case see [18].

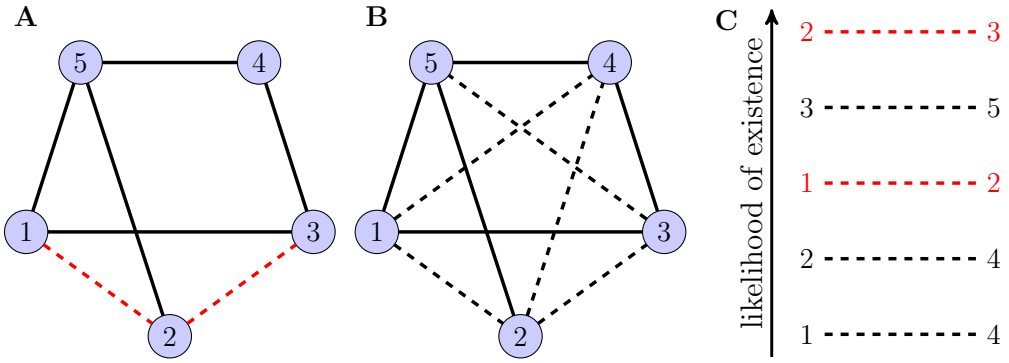

Figure 4.1: The Link Prediction Problem. We have incomplete knowledge about our network and only observe the five nodes (blue) and five links between them (solid lines), but there are actually two further links (red dashed lines) that are unobserved (**A**). To predict which links are missing, we rank all possible links (dashed lines in **B**) by their likelihood of existence (**C**). Finding appropriate likelihoods that lead to a correct ranking of links is the task of the link prediction algorithm. In this case the link $(3, 5)$ is incorrectly ranked above the link $(1, 2)$.

Statistical Testing by Bootstrapping The most accurate option to evaluate the performance of the algorithms would be to use all of the available data, make new predictions and then test these predictions by experiments. This approach is for obvious reasons inefficient and not feasible. So instead of predicting new and so far experimentally unobserved links, we will remove a subset of links from our data at random (using it as the aforementioned link set E^P) and try to recover these removed links on the basis of the remaining network (*i.e.* using it as the set E^T). This practice explains the superscripts, as these sets are called the "probe" and "training" sets, respectively. An overview over the various link sets that play a role in the link prediction framework is given in Table 4.1. We are going to refer to this approach as the "bootstrapping approach". The removal of links will also be referred to as the "dropout" (of links). The fraction of dropped links relative to the amount of existing links is called "dropout rate" or "dropout fraction". It is also important to note that the prediction accuracy obtained by this practice merely represent a lower bound since actual unobserved links will be counted as negatives.

Table 4.1: Overview and description of link sets in the link prediction process tested by bootstrapping.

Link Set	Description
U	Universal set. All possible links between the existing nodes.
$E = E^T \cup E^P$, $E^T \cap E^P = \emptyset$	All existing links (observed or not) in the network.
E^T	Training set. This is the known information for the algorithm to make predictions.
E^P	Probe set. This information is used in the evaluation but not for predictions.
$U - E^T$	Predictions. To each of these links the algorithm will assign a likelihood.
$U - E = (U - E^T) - E^P$	Non-existing links. These links will be false positives if they are predicted.

Testing with Theoretical Data The bootstrapping approach is the classical way to evaluate the performance of link prediction algorithms. However, as we have two different datasets for the hydrogen atom that are of experimental and theoretical origin, we are able to use the theoretical data as a probe set for the experimental data and thus able to avoid performing a dropout of links. This practice will be discussed in Section 4.2.5.

Evaluation Metrics The methods that we will study produce different kinds of scores that can not be interpreted as probabilities and can thus not be directly compared to each other. We will quantify the performance of the algorithms, by quantifying how correct predictions are favoured over wrong predictions in the ranked prediction list. This is done by plotting the *receiver operating characteristic curve* (ROC curve) [13].

The ROC curve is created by plotting the true positive rate (TPR), also known as "sensitivity" or "hit-rate", against the false positive rate (FPR), also known as "fall-out" at various threshold settings [13]. In our case the threshold parameter is the rank of the last considered entry in the ordered list of predictions, *i.e.* the number of predicted links that we are willing to accept and incorporate into the network. We choose to compare the two characteristics TPR and FPR at each rank in the prediction list. A perfect classification of the links with neither false positives nor false negatives would mean that the ROC curve rises vertically from the origin to the coordinate $(0, 1)$, then continuing horizontally to $(1, 1)$, just as a unit step function [13]. In contrast, for a prediction list done by random guessing, the ROC curve would follow the diagonal line (barring minor fluctuations) from $(0, 0)$ to $(1, 1)$. From this it follows that points above the diagonal line represent a classification that

is better than random while points below represent predictions that are worse than random. However, being consistently worse than random is better than random if we were to invert the output our classifier [13]. To summarise: the ROC curve shows the ability of a probabilistic classifier (whose output are scores, *i.e.* numeric values) to rank the positive instances relative to the negative instances [13]. An example case of the ROC curve is depicted in Figure 4.2.

To evaluate the performance of the algorithm according to the whole prediction list, we use the *area under the receiver operating characteristic curve* (AUC) metric [30]. This compresses the two-dimensional information given by the ROC curve into a single scalar value. The AUC value can be obtained by calculating the integral of the ROC curve [13]. As the AUC value is a portion of the unit square it varies in value between zero and 1. However, we established that random guessing leads to a diagonal ROC curve that has an AUC value of 0.5. Thus, classifiers with AUC values below 0.5 are being applied incorrectly [13]. The AUC value is equivalent to the probability that the classifier will rank a randomly chosen correct prediction higher than a randomly chosen false prediction [13].

Considering the ROC curve allows for better understanding of the algorithms and how their performance changes with respect to the increase of the dropout fraction and the type of network that is considered (*i.e.* which atom or data type is investigated). If no fundamental differences in the form of the ROC curves is found, we can subsequently use the AUC statistic to give a more accessible criterion for which method performs best. However, one should keep in mind that in practice only the top entries of the prediction list will be of immediate value. For further details about the ROC and AUC statistics please refer to [13].

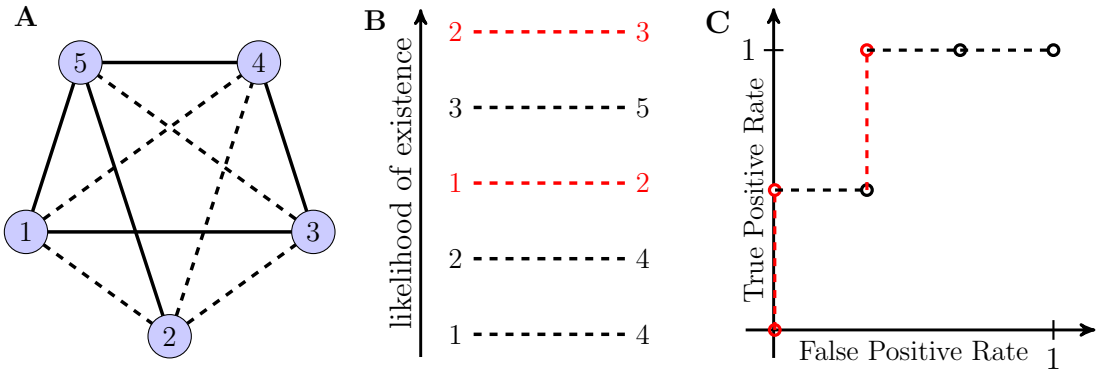

Figure 4.2: The construction of the *Receiver Operating Characteristic* (ROC) curve from a list of predictions. This figure illustrates how an ordered list of link predictions (**B**) for all unobserved links of a small network (**A**) is transformed into a ROC curve (**C**). Each point in the curve corresponds to the ratio of the sensitivity and fall-out among the predictions up to a particular rank, starting with rank zero. The first prediction is correct, hence we move up on the axis of the true positive rate by 1/2 (because we only have two links in the prediction set) to the coordinate $(0, \frac{1}{2})$. The link with the second highest likelihood of existence is not in the probe set E^P and hence a wrong prediction. As this link makes up a third of the overall number of false positives, the next coordinate is $(\frac{1}{3}, \frac{1}{2})$. Continuing like this for each rank in the prediction list and subsequently connecting all coordinate points creates the ROC curve.

4.2 Prediction of Atomic Transitions

4.2.1 Similarity-based Algorithms

A simple way to propose new links is by the assumption that similar nodes should be connected. In social networks for example this is motivated by the notion that people who have the same interests would make good friends. The question at hand is then how to define and quantify such a similarity for each pair of nodes i and j in a network [30]. A natural way to do this would be by the amount of mutual attributes - like mutual interests or hobbies - in social networks. Because we restrained from using such additional information in spectroscopic networks, we have to work solely on the basis of the structural data of the network, *i.e.* we need to focus on "structural similarity".

There are two fundamental concepts of structural similarity in networks: *structural equivalence* and *regular equivalence* [35, 7.12]. The former assumes that a link between two nodes indicates similarity between them, *i.e.* the basic idea is that nodes

are more similar if they have mutual neighbours [30]. The regular equivalence suggests that two nodes are similar if their neighbours themselves are similar [30][35, 7.12]. This can be interpreted as the network looking similar from the respective nodes perspective or as the nodes having similar 'roles' in the network [20]. The simplest way to predict links is therefore to calculate a score - or index - by a mathematical formula which somehow depends on the (immediate) neighbours of the nodes for each possible pair of nodes, and then rank the possible links by this score. The classification of similarity indices according to structural or regular equivalence will only hold up for extreme cases and the typical structure-based similarity index can be imagined being on a spectrum between the two extremes, incorporating both notions to some extent.

The link prediction with indices typically involves only simple algebra for each pair of nodes so that the computational effort typically scales with $\mathcal{O}(2N^2)$, where N is again the number of nodes in the network. The performance of these indices can therefore be seen as a "benchmark" for more sophisticated algorithms. Any algorithm that is substantially more complex should also perform substantially better. However, we do not care about the exact scaling of the link prediction algorithms, but we will judge them in terms of feasibility, *i.e.* we will focus on whether the application of an algorithm on the larger spectroscopic networks can be carried out within reasonable time scales.

Jaccard index

The *Jaccard index* is also known as 'Intersection over Union' (IoU) and is for example also used in the field of computer vision, where it is an evaluation metric for object detection algorithms. The index is defined as [30]

$$s_{ij}^{JC} = \frac{\Gamma(i) \cap \Gamma(j)}{\Gamma(i) \cup \Gamma(j)}, \tag{4.1}$$

where in the context of networks the sets $\Gamma(i)$ and $\Gamma(j)$ are the sets of neighbouring nodes of node i and j respectively. The motivation behind this index is quite intuitive, especially in the case of social networks: the more friends two persons share in comparison to their total number of friends, the more likely it is that they are also friends.

Performance Figure 4.6 depicts the performance of the link prediction based on the Jaccard index. The different subfigures A to C show the cases of different dropout fractions, with respectively 10 %, 30 % and 50 % of the initial links of the networks being used as the probe set E^P, whereas the remaining links were assigned to the training set E^T. The five coloured curves in each subfigure are the ROC curves for the respective networks as stated in the legend. Each curve is the result of vertical averaging over the ROC curves of 100 distinct simulations in which the links for the probe set E^P were each time chosen uniformly at random (for details regarding the vertical averaging see [13]). The shaded area above and under a curve indicates the standard error of the mean. The AUC values for all ROC curve are summarised in in Table 4.3 at the end of Section 4.2.

In some cases the ROC curves do not end in the point $(1, 1)$. This happens when the random assignment of links to the probe set left the remaining training network unconnected and these links are not considered in the prediction list. This primarily happens for high dropout fractions in small networks and is of low relevance for the obtained results.

The most prominent detail in this figure is the near-perfect performance of the algorithm in the H_{JB} network with an AUC of 0.996. In contrast, the performances in the H_{NIST} and the iron networks are quasi-random with values of 0.554 and 0.500, respectively and the performances in helium and carbon is even worse than random with AUC values < 0.4. We established the differences in structure between H_{JB} and the other networks in Section 2.2.2, but from this experiment alone it is hard to pinpoint what particular network properties influenced this algorithm. A possible explanation could be the higher network density that this network exhibits in contrast to the others, but it seems unlikely that this alone can explain near perfect result. Therefore an additional reason for this difference in performance could be that the theoretical hydrogen network also includes higher order atomic transitions and can be considered a "regular" network up to a certain type of transition (as explained in 2.2.2) that exhibits a high transitivity. The algorithm seems to easily find the gaps in this regularity induced by the random dropout. For the other networks one can imagine that the trend to bipartivity in the networks due to *Laporte's* selection rule for electric dipole lines is not captured by the Jaccard index. Rather than being likely to be connected when having common neighbours two nodes in the

spectroscopic networks seem to be more likely to be connected when the opposite is true. That this inverse would a better metric for the likelihood of links can be seen in the curves of helium, carbon and iron which all show a performance for the first entries which is substantially worse than random. If we were to invert the Jaccard index, this would lead to a higher performance at least for the top percentage of the predictions, especially in the case of carbon.

In the H_{NIST} network the algorithm leads to a performance that is in some parts worse and in some parts better than prediction by random guessing. The amount by which it excels a random prediction is higher than the amount by which it falls short in the other part, hence the AUC of 0.55. However, as in the other NIST-based networks, the latter takes place for the first entries in the ordered prediction list, so that an inverse Jaccard index would also be beneficial in this case.

When we compare the performance with respect to the dropout rates we notice that the AUC for H_{JB} drops only by 0.001 when we increase the dropout rate by 40 percentage points to 50 %. For H_{NIST} we see a slight incline by 0.024 and similarly small values for the remaining networks. One should be aware, that such an incline would however count as a loss in predictive power if we were to invert the index for these networks. These numbers show a robustness of either the algorithm or the networks structure against a dropout of links. The fact that the change is significantly smaller in H_{JB} suggests that the robustness stems from the network structure, as this network has a higher density.

Overall we can state that neither the Jaccard index nor an inverted version of it is suited for tackling the link prediction problem in spectroscopic networks. Although the performance is excellent for H_{JB}, one generally considers experimental data and the performance on such networks is not satisfying in terms of their AUC value and because of the strong dependence of the prediction accuracy on the threshold parameter (the rank of the last included entry in the prediction list). The huge disparity between the performances for the H_{JB} network and the NIST-based networks shows the lack of generality of this link prediction method. However, this was expected as the earlier comparison of these networks showed that this network is fundamentally different in composition. There is also an apparent disparity in performances within the group of NIST-based networks. The predictive power of this classifier

is weak for the iron network. The results suggest that the predictive power of this method depends on the complexity of the underlying atomic system. However, this dependence is not strictly monotonic, since the predictive power for carbon is higher than for helium. There are other undetermined factors that apparently influence the performance.

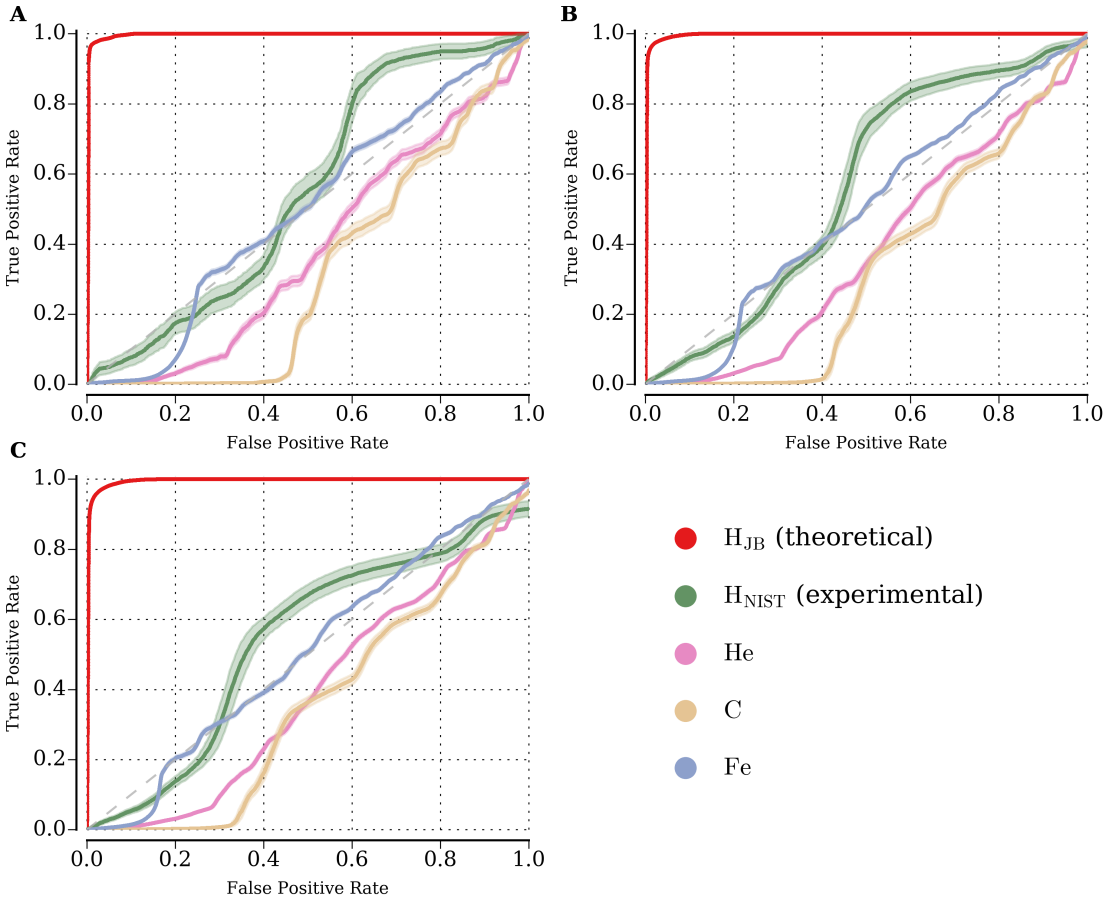

Figure 4.3: *Receiver operating characteristic* (ROC) curves for the link prediction carried out with the **Jaccard index** on the five example networks (indicated by colours) at the different dropout rates 10 %, 30 % and 50 % (figures **A** to **C**). In order to evaluate the quality of the link prediction, we delete a certain fraction of edges (dropout rate) and try to recover these deleted edges on the basis of the remaining network. The ROC curve shows the values of the true positive rate (TPR) against the false positive rate (FPR) at various fractions of the prediction list. The value of the area under the ROC curve (AUC) statistic for each curve can be found in Table 4.3. The graphs are discussed in more detail in the main text.

Adamic-Adar index

The basic assumption of the Adamic-Adar index is to measure similarity by the number of common neighbours that the nodes i and j have. This is analogue to the idea that two people are more likely to be friends if they have many friends in common. However, in this index the nodes with a higher degree are attributed a lower weight in the count [30]. In social networks this can be understood intuitively by the notion that I rather know the friends of a person with few friends than the friends of an immensely popular person. The weighting is done via the log-function, so that [30]

$$s_{ij}^{AA} = \sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log k_z}, \tag{4.2}$$

where k_z is the degree of node z.

Performance If we consider the first dropout rate, the performance of this index is overall similar to the on the Jaccard index, again the performance for the theoretical network H_{JB} is very high with an AUC value of 0.996, which is in accordance to the AUC of the Jaccard index up to two digits after the decimal point. The performance for H_{NIST} is considerably higher than for the Jaccard index with a value of 0.749. The ROC curves for helium and carbon are again over their entire range below the diagonal line. Therefore one would again receive better predictions if one had taken the inverse of the Adamic-Adar index in these cases. This is also true for the very first few predictions for the iron network, but above a false positive rate of 0.2, the curve is mostly following the bisecting line with some minor fluctuations around it, leading to an AUC of 0.515. The outcomes can be understood by considering the size and complexity of the systems: while helium and carbon have small atomic numbers of 2 and 6, iron is a much more complex system.

Substantial changes with increasing dropout are only seen for the H_{NIST}: the AUC of the Adamic-Adar index in H_{NIST} decreases by 0.088 to 0.661. For the other networks, the change in value is never greater than 0.4. As this is the smallest network, this suggests that especially small networks are susceptible to dropout of links.

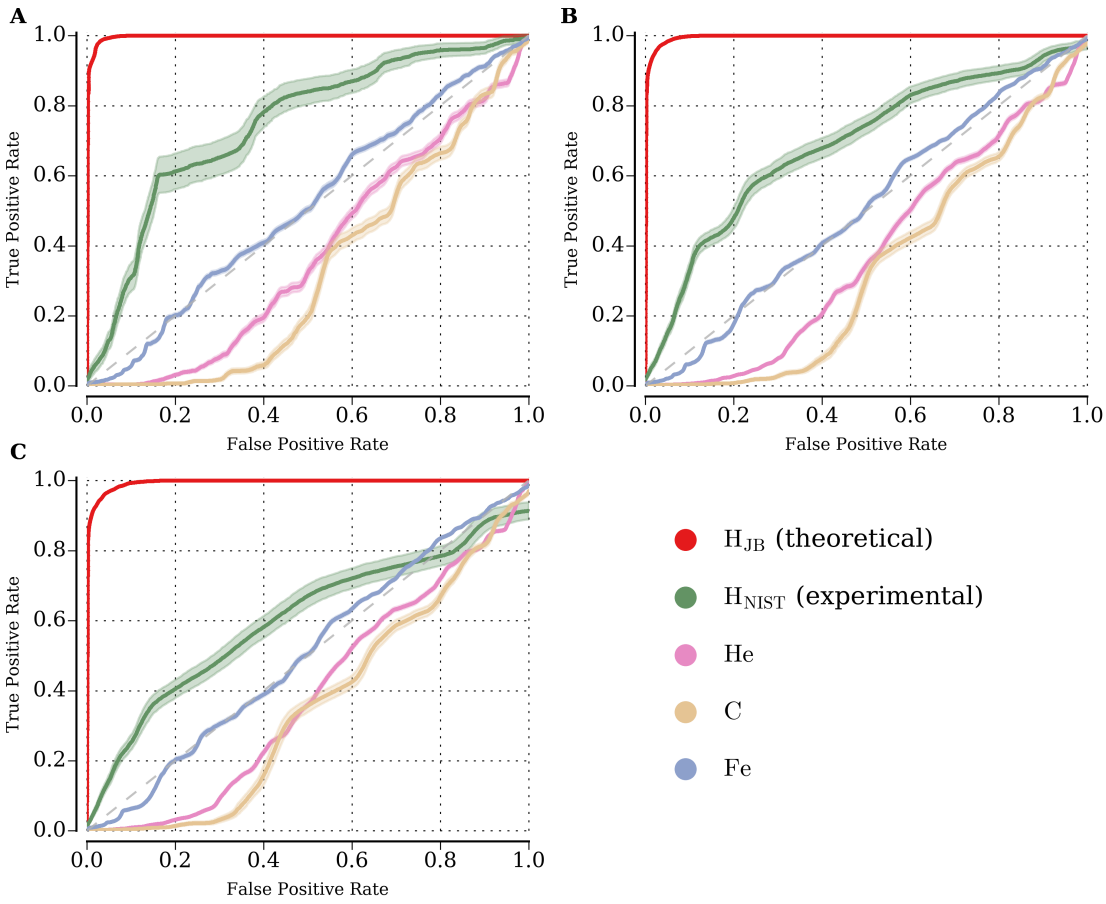

Figure 4.4: *Receiver operating characteristic* (ROC) curves for the link prediction carried out with the **Adamic-Adar index** on the five example networks (indicated by colours) at the different dropout rates 10 %, 30 % and 50 % (figures **A** to **C**). For a detailed description see caption of Figure 4.3.

Preferential Attachment index

We introduced the notion of preferential attachment in the context of scale-free networks in Section 2.2.1. In [2] it is discussed how a evolving network, where the addition of a new link to nodes i and j is dependent on the product of their degrees $k_i \cdot k_j$, will exhibit a scale-free degree distribution. A similarity index based on the idea of this mechanism is then defined as [30]

$$s_{ij}^{PA} = k_i \cdot k_j. \tag{4.3}$$

Since at least some of the spectroscopic networks have shown qualities of scale-free networks, it is worthwhile to study a similarity notion based on that concept.

Also, because for this index no information about the nodes neighbours is needed, it has the lowest computational complexity of all the algorithms considered: instead of scaling with $\mathcal{O}(2n^2)$ as the other similarity indices in this section do, the computationally cost only grows with $\mathcal{O}(2n)$ when the network size is increased [46]. However, these differences do not seem to be important for the network sizes of typical spectroscopic networks as calculations for both types are executed in a matter of seconds even for the larger spectroscopic networks on a local machine.

Performance This algorithm is the first for which its prediction are consistently better than random predictions for all considered networks. The best performance is again achieved for the H_{JB} network, although the ROC curve in Figure 4.5 has a sharp kink at a true positive rate of about 0.85, after which almost only wrong predictions follow, before the curve returns to a quasi-random behaviour for the last few predictions. The AUC value of this curve is 0.865. We could not find an intuitive explanation for this unusual behaviour. The difference to the performance in the remaining networks is still substantial, but not as great as for the previous two methods: the method performs worst for helium at an AUC value of 0.691, while for the others the performance is very alike with AUC values of 0.820, 0.852 and 0.827 2 for H_{NIST}, carbon and iron respectively (and all being measured for a dropout of 10 %). The performance does again not strictly correlate to the complexity of the atomic systems. This suggests that other structural factors also play a role, but a more extensive investigation would be needed to explain why the Preferential Attachment index performs worse in the helium network than in the carbon network.

There are only small declines in performance (all below an change in AUC of 0.03) when increasing the dropout rate from 10 % to 50 %. It is also notable that the kink in the ROC curves for H_{JB} does not change its position on the TPR axis for different dropout rates.

Due to its consistently better-than-random prediction behaviour, the Preferential Attachment index seems to be better suited for link prediction in spectroscopic networks than the previous ones.

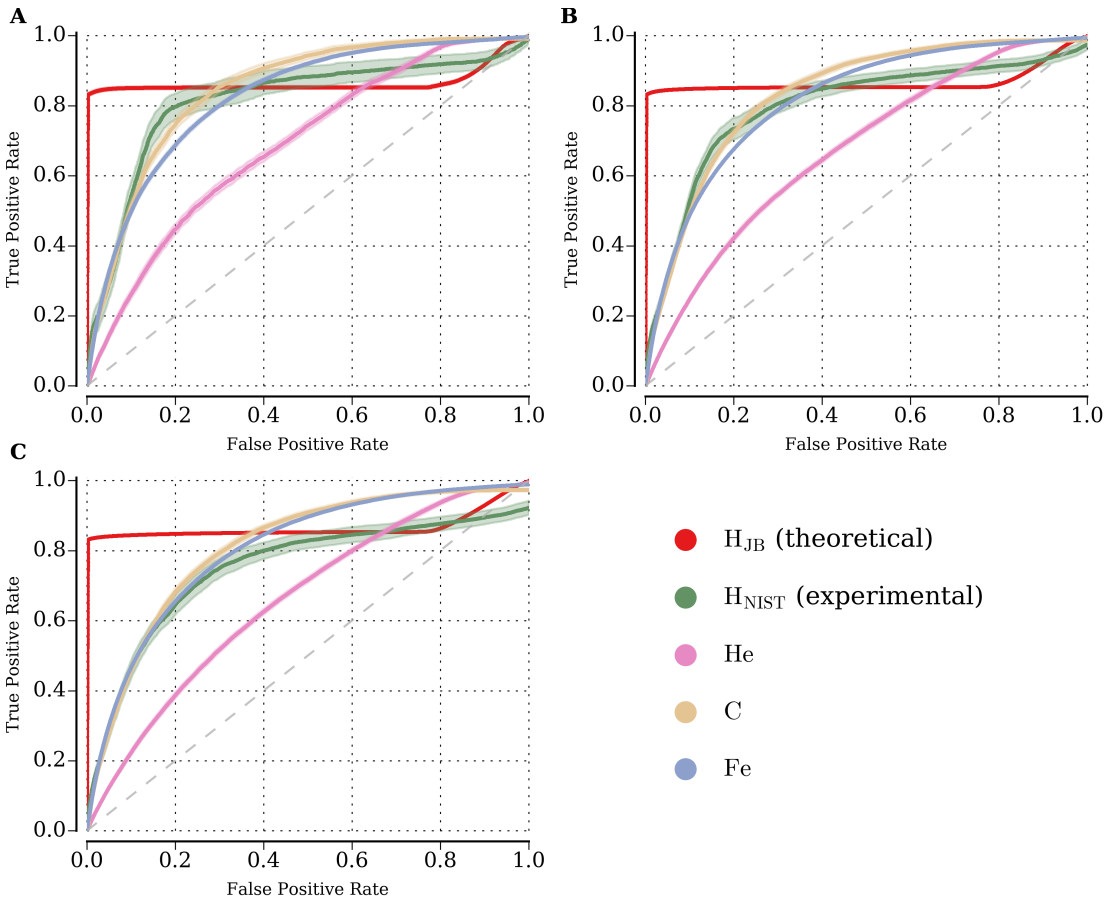

Figure 4.5: *Receiver operating characteristic* (ROC) curves for the link prediction carried out with the **Preferential Attachment index** on the five example networks (indicated by colours) at the different dropout rates 10 %, 30 % and 50 % (figures **A** to **C**). For a detailed description see caption of Figure 4.3.

Resource Allocation index

The *Resource Allocation index* is motivated by the resource allocation dynamics on networks, where a resource is transmitted from node i to node j via their common neighbours. The similarity is defined by the amount of resource j receives from j [48],

$$s_{ij}^{RA} = \sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{1}{k_z}. \tag{4.4}$$

Although motivated from a dynamical process, it is similar to the Adamic-Adar index, with an even higher suppression of the contribution from high-degree nodes to the score [30].

In [48] the authors compared a number of local similarity-based algorithms on a variety of networks. For that sample of networks, the Resource Allocation index performed best in each case.

Performance The mathematical similarity to the Adamic-Adar index becomes apparent in the results. While again this method performs excellent for the theoretical hydrogen network H_{JB} with an AUC of about 0.996, its predictions are also substantially better than random predictions for the hydrogen network created from NIST data, for which the performance has a value of 0.762 and for any other network it performs near-random or worse than a random prediction: the AUC values of helium, carbon and iron are 0.385, 0.332 and 0.516 2 respectively for a dropout rate of 10 %. All these AUC values agree with those of the Adamic-Adar index within the error bounds, suggesting that the different weighting functions of these indices does not influence their predictive power in spectroscopic networks. The Resource Allocation index has therefore the same limitations as the Adamic-Adar index that were already discussed in detail.

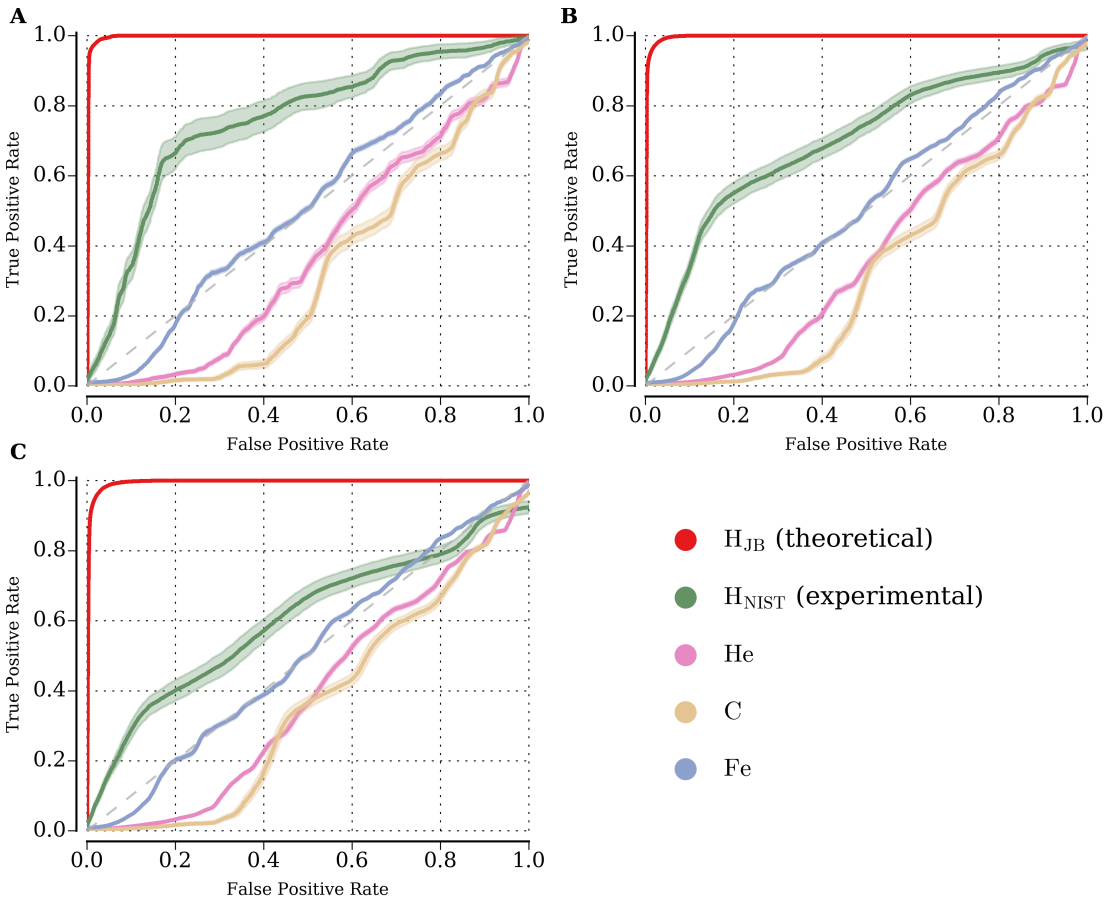

Figure 4.6: *Receiver operating characteristic* (ROC) curves for the link prediction carried out with the **Resource Allocation index** on the five example networks (indicated by colours) at the different dropout rates 10 %, 30 % and 50 % (figures **A** to **C**). For a detailed description see caption of Figure 4.3.

Conclusion

The comparison of local similarity-based indices by Zhou *et al.* in [48] yielded the Resource Allocation index (which was introduced in said publication) as the best performing index (measured by AUC value), followed by the Adamic-Adar index as the second best and leaving the Preferential Attachment index as the index with the worst predictions of all 10 considered indices. In contrast to that result, there was no performance difference between the Resource Allocation index and the Adamic-Adar index for the spectroscopic networks in this work, and - more surprising - the Preferential Attachment index performed best for all spectroscopic networks but the theoretical hydrogen network.

However the results also show that none of the similarity-based methods were able to adapt to the differences in the network structure of the different spectroscopic networks. Even the Preferential Attachment index leads to different forms of the ROC curves for the networks of different data types. It is not a necessity, that the algorithms perform equally well for both types of data but it suggests a lack of generality if they do not.

4.2.2 Hierarchical Structure Method

While the previously described attempts at link prediction used the degree of nodes as the main quantity for the determination of likelihoods, the following approach is more holistic as it seeks to describe the structure of the whole network by means of a *dendrogram*, also called a *hierarchical random graph*, before deriving link scores.

The central assumption of the method that Clauset *et al.* proposed in [7] is that real-world networks are hierarchically organised, a claim founded on empirical evidence of previous studies [43]. Although such an underlying hierarchical structure might not be the case in every network, the discussion of the community structure of spectroscopic networks in Chapter 3 justifies an approach based on this assumption for this network type.

The authors proposed both a general technique to uncover this hierarchical structure and an algorithm to predict missing links based on this structure. This underlying hierarchical network structure is mathematically modelled by the means of a dendrogram D. A dendrogram D is a binary tree whose leaves are the nodes of our network graph G [7]. The way the the leaves are merged into groups depends on the network structure of G, but is not uniquely given. For each merger (represented by internal nodes of the tree D) there is an associated probability in the dendrogram (which is 1 if all the leaves - *i.e.* nodes in G - that are being merged are connected in the network graph G). The network structure of G is thus encoded by the hierarchical tree structure D and the probabilities associated to internal nodes. In this model it is also possible to define the degree of relatedness of two leaves in D. The likelihood for the existence of a connection in G is thus given by this relatedness of two nodes within the found hierarchical network structure [7]. Because the representation of the network structure in terms of a dendrograms is not unique and

there are often several different plausible dendrograms of roughly equal likelihood. From this ensemble of possible hierarchies, the method obtains its predictions by a weighted sampling and thus the method is supposed to avoid over-fitting the data [7]. A more detailed explanation of the method can be found in Appendix A, [30] and [7].

A promise of this method is that it not only captures the hierarchical structure with respect to the notion of assortativity, but is as well able to model disassortative structure. This is done via the probabilities of the internal nodes of a dendrogram, where a high probability relates to assortative structure and a low probability to disassortative structure [7]. This generality sets this method apart from the previous methods. However, an obvious shortcoming of such a maximum likelihood method is the computational cost due to the many intermediate step such as creating dendrograms, evaluating their likelihood, merging the multitude of plausible dendrograms and evaluating the likelihoods for the single links. It is possible to use these methods on networks with a few thousand nodes, but big networks cannot be studied in reasonable time on a typical CPU [30].

Performance The results in Figure 4.7 have been obtained as for the previous similarity-based methods. Again we have the case that the performance of the method is substantially worse in one network than in the others. Again it is not only worse, but worse than random guessing, suggesting that in this network it would also be advisable to follow the inverse of the proposed likelihood. With an AUC value of about 0.977 this method performs slightly worse compared to the JC, AA and RA indices, but better than the PA index for the H_{JB} network. For the H_{NIST}, helium and carbon networks, it performs better than all previous considered methods with AUC values of 0.892, 0.863 and 0.894 respectively. For the Iron network, the prediction is with an AUC of 0.377 worse than random guessing.

The small deviation of the ROC curve in the iron network form the diagonal suggests that the network does not exhibit a strong hierarchical structure. A possible explanation is that the selection rules do not apply as strictly in this atomic system. Because of the higher number of electrons in this system and the interactions between them the previously discussed quantum numbers are not appropriate any more as their operators do not commute with the Hamiltonian. As a greater set

of quantum numbers is needed to uniquely identify the eigenstates of the system the network loses the concise structure that was present in the networks of smaller quantum systems. In contrast the performance for the H_{JB} network yields a high AUC value and is similar to the one the similarity-based algorithms, although the method is not solely based on local information. This suggests that for this network a hierarchical structure could be found. This is notable, because, as we explained in 2.2.2, the different transitions types are treated equally.

The HRG method is outperforming the Preferential Attachment method for every network but the iron network based on the AUC statistic. While the Preferential Attachment index was the least time-consuming of all methods, the HRG method includes many more intermediate steps, and can take up to several days of computing time on the CPU of a local machine.

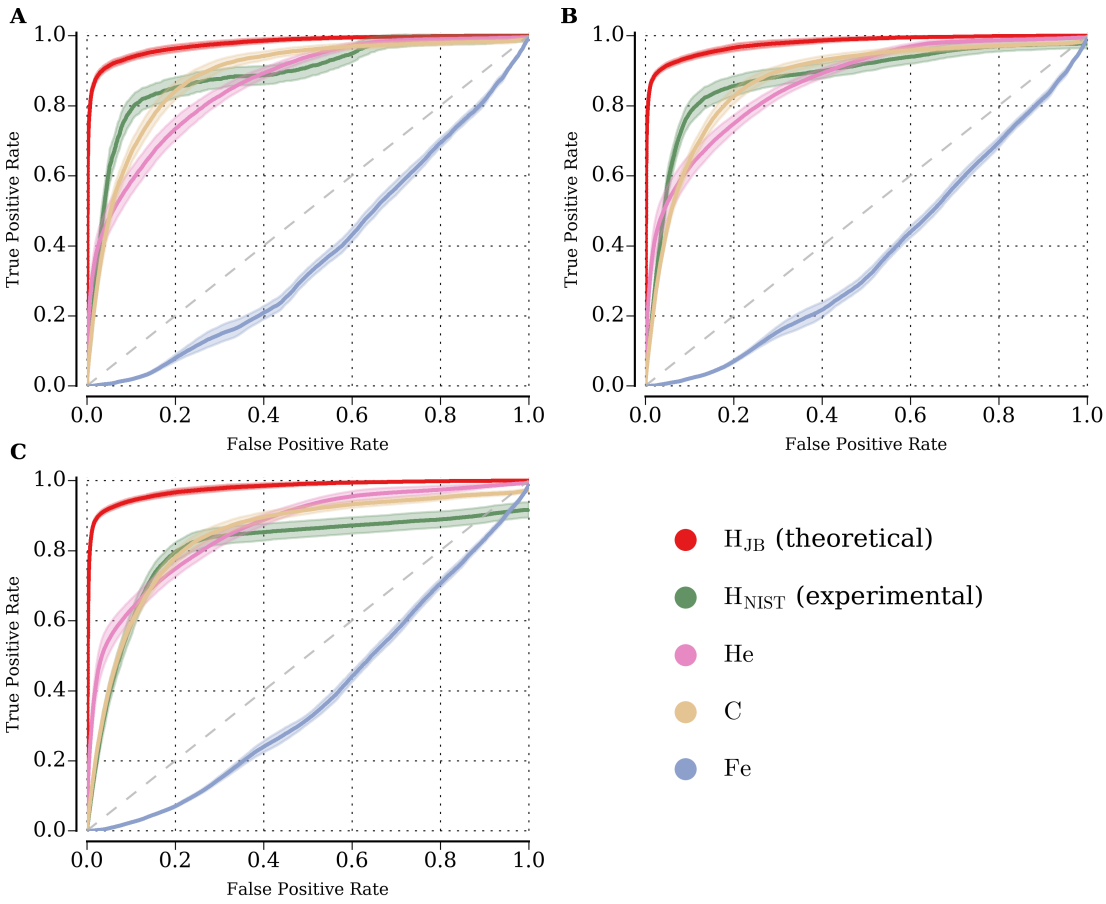

Figure 4.7: *Receiver operating characteristic* (ROC) curves for the link prediction carried out with the **hierarchical random graph method** (HRG) on the five example networks (indicated by colours) at the different dropout rates 10 %, 30 % and 50 % (figures **A** to **C**). For a detailed description see caption of Figure 4.3.

4.2.3 Nested Stochastic Block Model Method

In Chapter 3 it was established that one can use the stochastic block model method to infer the community structure of a network. It is also possible to use this approach to assign a score to the existence of a link in the network. The basic approach to extend this model to be able to predict links is to add a possible links to the network and calculate the likelihood of this change within the nSBM framework (*i.e.* under consideration of the community partitioning). Doing this for each possible link and normalising each likelihood by the sum of all likelihoods, we generate the prediction list [45]. Due to this repeated calculation of the network entropy in the derivation of the liklihood, this method is even more time consuming than the previous maximum

likelihood method based on hierarchical random graphs and thus is not suited for networks of the order of 10^3 nodes, as the computation time on a single CPU gets infeasible.

Performance Figure 4.8 shows that the performance of this method is consistently high for all considered spectroscopic networks. Although the AUC value for the H_{JB} network is second lowest of all previous methods, it is still as high as 0.908. In contrast to the HRG method, the nSBM algorithm held its promise of generality as all AUC values are in a interval $[0.896, 0.9626]$ for the lowest dropout rate. The highest AUC value is yielded in the iron network, which sets this method apart from the remaining methods. However, because of the high computational cost, this curve shows the result of a single simulation of the method and not an average over 100 as for the other networks.

As we discussed in Chapter 3 there was only a weak correlation between the communities found by the nSBM method and the quantum numbers L, J and S in the H_{JB} and iron network. The high AUC values for these networks suggest now that the algorithm was able to find other features than quantum numbers by which the network is structured. Overall, the results of this method are superior in almost all cases to both the HRG and the PA method. However, due to the computational cost it should only be used for small systems.

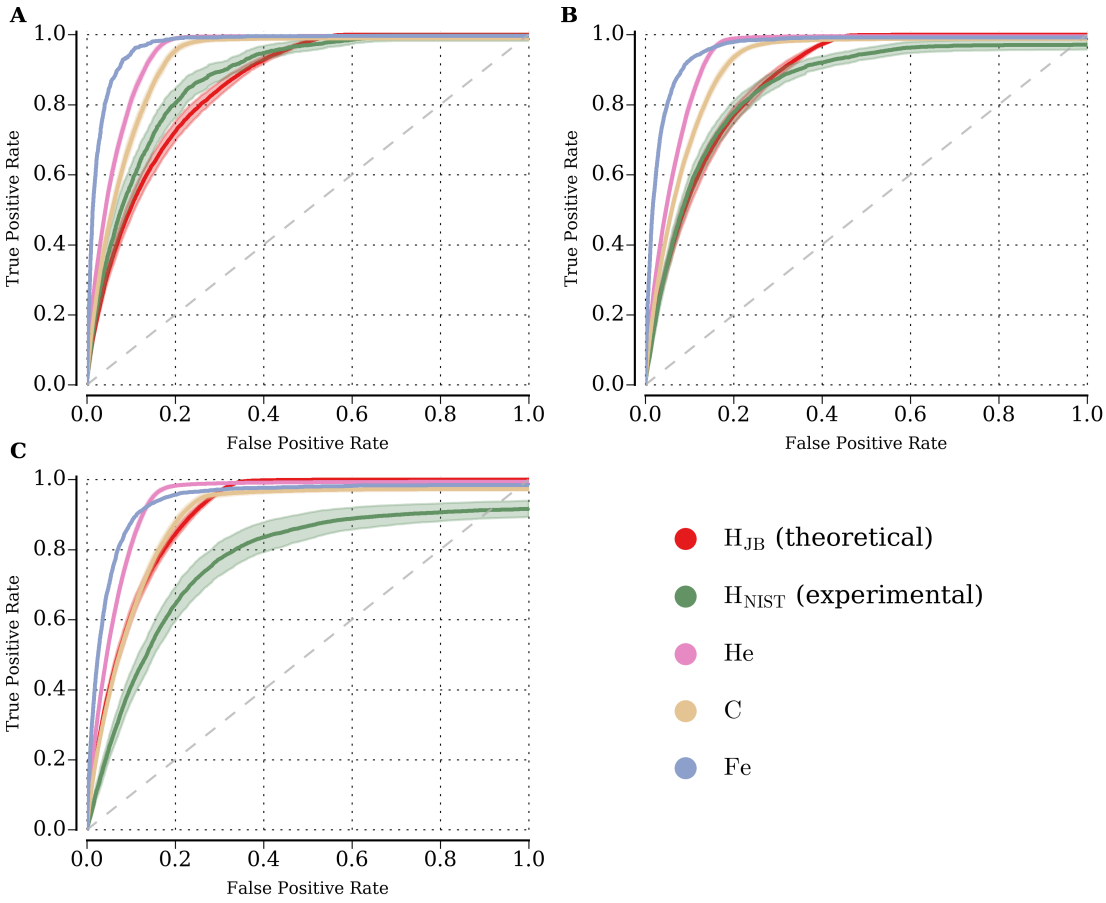

Figure 4.8: *Receiver operating characteristic* (ROC) curves for the link prediction carried out with the **nested stochastic block model method** (nSBM) on the five example networks (indicated by colours) at the different dropout rates 10 %, 30 % and 50 % (figures **A** to **C**). Due to high computational cost, ROC curve of iron is the result of a single execution of the algorithm and not an average over 100 runs. For a detailed description see caption of Figure 4.3.

4.2.4 Structural Perturbation Method

The following link prediction method is based on the work of Lü *et al.* in [29] and is in opposition to the previous methods neither a similarity-based method nor a maximum likelihood method. Rather, they assume that the regularity of a network is reflected in the "consistency of structural features" before and after a perturbation of the adjacency matrix [29]. This means that they do not assume any a priori organising principle of the network.

The authors propose a *structural consistency index* σ_c that is supposed to reflect the "inherent link predictability" of a network. It is obtained by a first-order matrix perturbation of the adjacency matrix and the subsequent changes of eigenvalues; for further details see Appendix B and [29]. The motivation behind the notion of link predictability is that the structure of real-world networks is driven by both regular and irregular factors. The authors state that only the regular factors can be explained by mechanistic models - such as the preferential attachment process - and therefore set out to estimate to what degree a network follows such explicable behaviour [29]. The authors back this idea up by showing the correlation between the structural consistency and link prediction accuracy measured by the precision metric [29].

We instantly notice that their assumption of regular and irregular factors is not true in the case of spectroscopic networks: the composition of the energy levels and the optical transitions between them is strictly guided by deterministic laws of nature. However, one could argue that highly complex processes - such as spin-orbit coupling - that make systems with more electrons deviate from the simplicity of the hydrogen atom, cannot be captured by those simple mechanistic models and will therefore appear as irregular factors from their point of view. That being said the calculations in this algorithm are based on eigendecomposition and perturbation of the adjacency matrix and similar to the calculations of the first-order perturbation of the Hamiltonian in quantum mechanics.

In Table 4.2 we listed the values of the structural consistencies for the spectroscopic networks. The values are in a similar range as the values of the example networks in [29], ranging from 0.31 to 0.69, while the examples were ranging from 0.22 to 0.71. The results suggest an interpretation that we have already given in the discussion of

previous methods: the predictability seems to decrease with increasing complexity of the system, with the H_{NIST} network being an outlier due to its small size. The high value for H_{JB} is most likely due to the density and the high transitivity of this system. Because of these properties, any missing links can be found by the gaps they leave in this highly regular system.

To not only estimate a predictability of a network, but to actually predict links in it, the entries of the perturbed adjacency matrix that was used in the course of calculating the structural consistency can also be understood as the likelihood of existence for the missing links.

In contrast to the maximum likelihood methods, this method has a substantially lower computational cost. Even on the biggest networks that were considered in this project this algorithm could be executed within the order of minutes.

Table 4.2: *Structural consistency* σ_C of the respective spectroscopic networks. This metric was proposed by Lü *et al.* in [29] and indicates the link predictability of a network as it estimates the extent to which the organisation of a network is explicable. It varies in value between a minimum of 0 and a maximal value of 1. The uncertainty is given in terms of the standard deviation.

Network	Structural Consistency σ_c
H_{NIST}	0.52(5)
H_{JB}	0.933(1)
He	0.69(2)
C	0.63(3)
Fe	0.31(1)

Performance In Table 4.3 we see that in six of the 15 considered instances the *structural perturbation method* (SPM) is the top performing method in terms of the AUC value; in two further cases its mean is within the error bound of the top performing method. Only for the iron network it is consistently outperformed by the nSBM method. In only four instances the AUC value of the SPM is below 0.9. That the three worst results of this method are the predictions for the H_{NIST} network corresponds to the fact that this network also has a low structural consistency

index. It also turned out to be the network for which the prediction accuracies were most susceptible to an increase of the dropout. Nonetheless, an AUC value of 0.84 at a dropout fraction of 10 % suggests that the predictions by the SPM are still meaningful even in this network. It is also notable that the method does not show a dependence on the quantum mechanic complexity of the network (*i.e.* number of electrons in the atomic system), since the AUC values in the iron network are as well above 0.9. However this result is somewhat curious as the structural consistency for iron has a value of only about 0.3, which is the lowest value of all considered networks.

The results in Table 4.3 also suggest that the method is not as robust against an increase in the dropout fraction as the other methods, especially when compared to the maximum likelihood methods. This could be attributed to the fact that the method is not modelling the organising principle of the network and is therefore more dependent on the amount of data. Also the method itself is based on a bootstrapping approach for the creation of a perturbation (for a detailed explanation see appendix B or [29]) and hence the amount of data that can used for making predictions is even smaller for this method.

Overall this method performs consistently high, while also having a low computational cost, *i.e.* its application is feasible for the network sizes considered in this project.

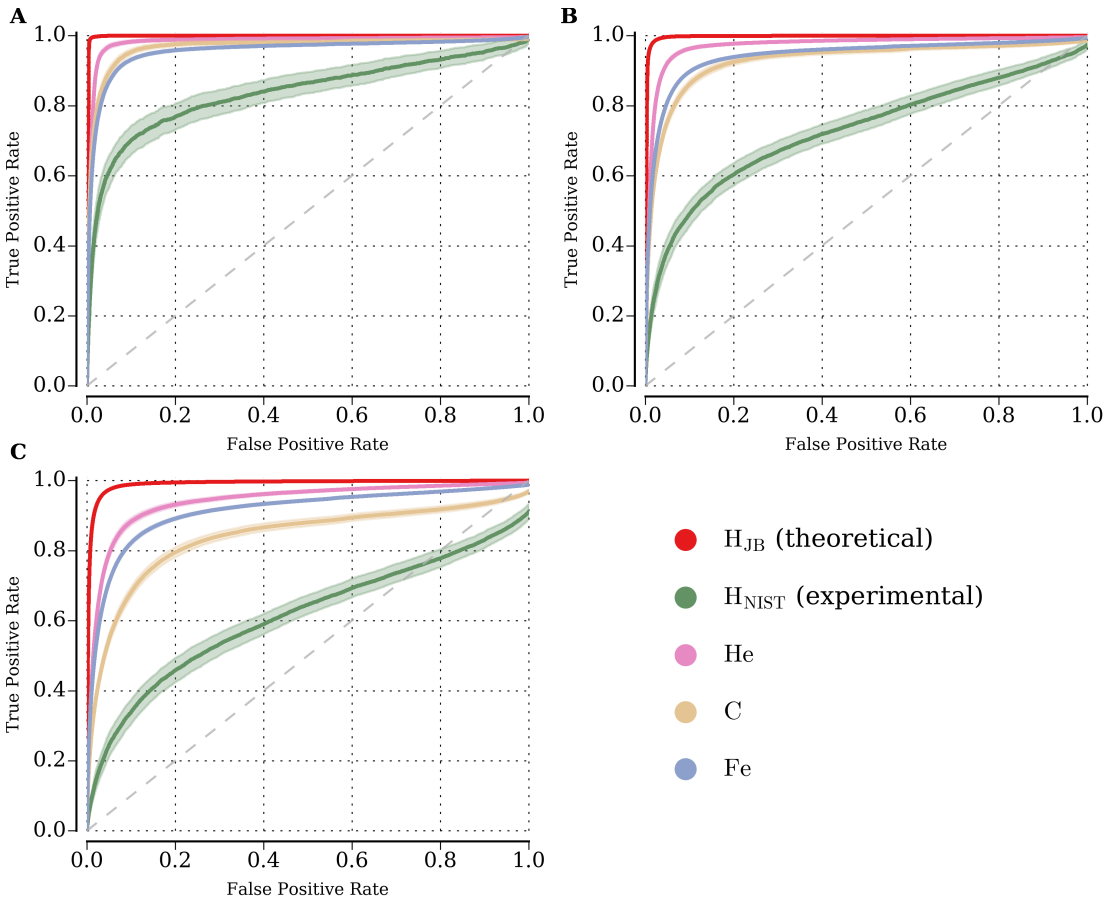

Figure 4.9: *Receiver operating characteristic* (ROC) curves for the link prediction carried out with the **structural perturbation method** (SPM) on the five example networks (indicated by colours) at the different dropout rates 10 %, 30 % and 50 % (figures **A** to **C**). For a detailed description see caption of Figure 4.3.

4.2.5 Evaluation with Theoretical Data

So far the results were validated by deleting edges at random and measuring how well the link prediction method could recover them. This was statistical testing by bootstrapping. In contrast to that practice, we now use a training set E^T that is entirely made up of experimental data as the basis for the predictions and a probe set E^P that is made up of theoretical data.

We will not consider all combinations of methods and networks as in the previous section, but since we established the high accuracy of the structural perturbation method, we will only use that method. Also, we will only consider the helium

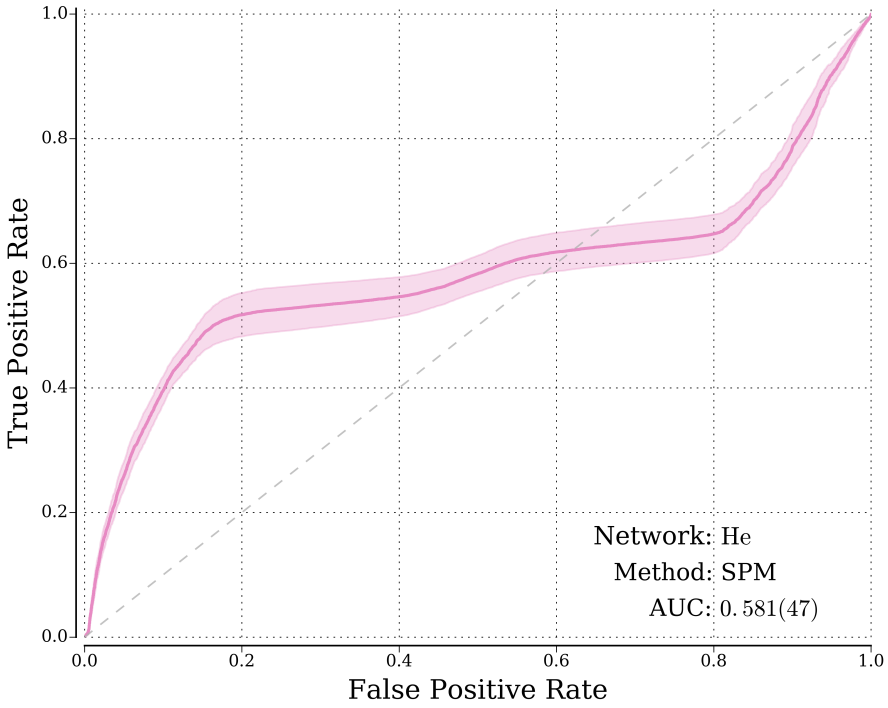

Figure 4.10: *Receiver operating characteristic* (ROC) curves for the link prediction carried out with the **structural perturbation method** (SPM) on the basis of the experimental data in the **helium** network. Predicted links are counted towards the true positives if they are part of the theoretically obtained data of the helium atom in the *ASD* data base [25]. The AUC value of this ROC curve is 0.581(47).

network, as for this data set it is possible to separate the experimental from the theoretical data by means of meta data labels.

Figure 4.10 shows the ROC curve of the evaluation of these predictions. The curve shows three parts with different behaviour: after a steep initial rise, the curve flattens after a FPR of about 0.2 and a TPR of 0.5, before rising again steeply in the last part after a FPR of 0.8. This suggests that different types of links exist for which the predictive power of the SPM differs. However it is not clear what to what feature this difference corresponds. The different types of transitions can be ruled out, as the data mostly consists of electric dipole transitions (see Table 2.1). It is also possible that there are effects due to the small size of the network that only features links which are the training set E^T. Yet even on the basis of this small experimental data set, the AUC metric yields a value of 0.581(47).

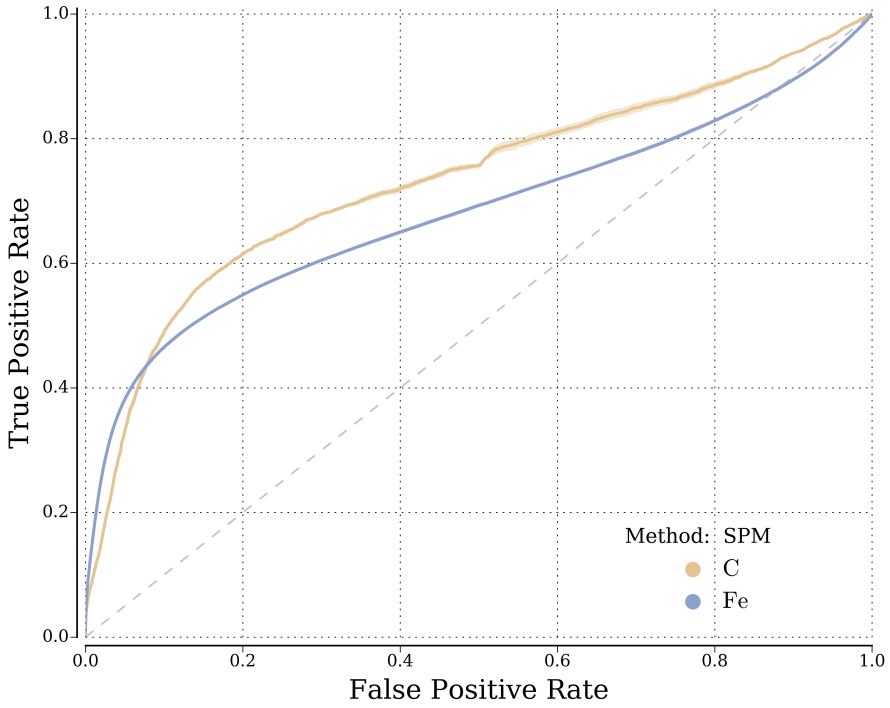

Figure 4.11: Agreement of link predictions with dipole selection rules. For this ROC curve predicted links are counted towards the true positives if they do not violate the selection rules for electric dipole lines. This is done with the predictions of the structural perturbation method (SPM) for the carbon and the iron network. The AUC value of the respective ROC curves are 0.724(22) and 0.674 7(25). See Figure 4.9 for the accuracy of these prediction.

Furthermore, we have checked to what extent the predictions by the structural perturbation method obey the selection rules for electric dipole transition. Figure 4.11 shows a ROC curve for the predictions by the SPM in which predictions are counted as positive instances if they are in accordance with the selection rules and as negative instances if not. This is done for the predictions for a dropout rate of 10 % in the carbon and iron networks (compare with Figure 4.9 **A**). We see in this graph that predicted links are more likely to adhere to the selection rules than not, as the curves for carbon and iron have AUC values of 0.724(22) and 0.674 7(25), respectively. However these values are not such that one could argue that the link prediction method merely learned these rules and applies them for prediction in a rigorous manner. Further investigation is needed to compare the prediction with the selection rules of other coupling regimes (such as jj-coupling and intermediate coupling) and other transition types (higher order electric and magnetic transitions).

4.3 Conclusion

We have demonstrated in this section that several link prediction methods were able to predict the existence of atomic transitions based only on the structural data of a spectroscopic network. This is true both for spectroscopic networks based on theoretically and experimentally obtained data, although not all methods could handle both types equally well.

The structural perturbation method [29] and the nested stochastic block model method [38] both performed well on all example networks, with the SPM outperforming the latter. The AUC values of its prediction were as high as ≈ 0.96 for the most complex considers atomic system, the iron network. Of the similarity-based indices, only the Preferential Attachment index [30] performed consistently better than random in all networks. The maximum likelihood methods overall performed better than the simplistic similarity-based indices, however one should be aware of their high computational cost that lead to problems with the largest considered networks. This trade-off between efficacy and computational could be resolved by using the SPM, as with this method the computing time in this project was of the order of minutes with a typical CPU.

In general the performances of link prediction algorithms provide evidence of the corresponding mechanisms at work for network organisation [29]. The performance of the PA index seems to confirm that the experimental data taking process of the spectral data could be modelled by a preferential attachment model, as we proposed in Chapter 2.2. The performances of the maximum likelihood methods indicate the existence of meaningful communities or even a hierarchical structure. For atoms with low atomic number we established to which quantum numbers this structure is correlated to. For more complex atoms with many electrons it is not clear, what the defining structural features are. These results together with the low predictive power of the similarity-based indices suggest a disassortative building principle in this type of networks, that is driven by the quantum mechanic selection rules. The SPM does not give meaningful clues about the building mechanism of the network as this method explicitly refrained from assuming any organising principle. That a variety of algorithms could successfully predict links indicates that there is not one network organisation principle at work like it is the case for purely mathematical

network models. Rather real networks exhibit a mixture of network principles and the composition of this mixture also changes with respect to how the data was obtained. The algorithms that were used in this work are the common state-of-the art methods, however, there are additional types of link prediction algorithms, such as probabilistic models [30], that were not covered in this work and might lead to further insights into the organising principles of spectroscopic networks or could simply increase the accuracy of the predictions.

The results also showed a robustness of the prediction accuracies against the dropout of links. It can not be deduced with certainty if this robustness stems from the applied methods or from an inherent robustness of the spectroscopic networks, but since the performance of several types of methods proved to be stable, it seems to be a property of the networks themselves.

Overall the results show that by using link prediction methods to spectroscopic networks we can make predictions of high accuracy about which unmeasured atomic transitions should exist. The advantages over a prediction on the basis of the selection rules is that this method automatically adapts to the complexity of the systems. *I.e.* even when the quantum numbers L, J and S are not appropriate any more due to failing of the LS-coupling, the data-driven methods adapt to this change of the network organising principle and continue predicting with high accuracy in regimes that are not as strictly adhering to a set of rules. To confirm this hypothesis in a rigorous manner a more extensive side-by-side comparison of the predictions on the basis of networks and those on the basis of the selection rules in should be carried out on the back of a ground truth data set. So far this has only been done for the selection rules of electronic dipole lines. Furthermore, by predicting transitions in terms of links, we are given a ranking of likelihoods and thus it is possible to find gaps in the data (like dipole transitions with high transition rate that have been overlooked in a measurement). A prediction solely on the basis of the selection rules could not provide this as efficiently.

Table 4.3: Link prediction accuracies of the different link prediction algorithms in the example networks for different dropout rates measured by *area under the ROC curve* (AUC). The AUC value is given portions of the unit square and varies in value between 0 and 1. Random guessing leads to a diagonal ROC curve and thus an AUC value of 0.5. Classifiers with AUC values below 0.5 are being applied incorrectly. The dropout rate indicates the fraction of edges that were initially deleted and subsequently used as the probe set E^P. The highest value for each combination of network and dropout rate is in boldface. The uncertainty is given in terms of the standard deviation. The values for the nSBM method in the iron networks are the result of a single run, while any other value is the average over 100 simulations. An overview of the networks can be found in Table 2.1. The methods are presented in the main text of this chapter.

Network	Method	Dropout Rate 10 %	30 %	50 %
H_{NIST}	JC	0.554(34)	0.578(26)	0.586(28)
	AA	0.749(50)	0.708(31)	0.661(32)
	PA	0.820(45)	0.818(23)	0.810(17)
	RA	0.762(46)	0.718(33)	0.655(36)
	HRG	0.892(33)	**0.900(25)**	**0.877(19)**
	nSBM	**0.896(22)**	0.888(17)	0.843(44)
	SPM	0.846(47)	0.747(39)	0.666(47)
H_{JB}	JC	0.99565(23)	0.99570(19)	0.99475(24)
	AA	0.99626(26)	0.99552(32)	0.99403(45)
	PA	0.8650(79)	0.8660(36)	0.8649(25)
	RA	0.99638(17)	0.99607(19)	**0.99485(27)**
	HRG	0.9769(86)	0.9775(82)	0.9770(80)
	nSBM	0.908(10)	0.9129(93)	0.9250(56)
	SPM	**0.99729(18)**	**0.99651(26)**	0.9910(11)
He	JC	0.385(15)	0.3856(87)	0.3924(67)
	AA	0.380(14)	0.3855(82)	0.3914(62)
	PA	0.691(15)	0.6792(86)	0.6619(86)
	RA	0.385(14)	0.3849(89)	0.3929(83)
	HRG	0.863(22)	0.870(17)	0.866(19)
	nSBM	0.9512(37)	0.9478(35)	**0.9469(37)**
	SPM	**0.9863(42)**	**0.9762(33)**	0.9460(79)
C	JC	0.321(16)	0.339(15)	0.367(14)
	AA	0.332(16)	0.343(14)	0.369(18)
	PA	0.852(13)	0.8449(76)	0.8313(74)
	RA	0.332(19)	0.346(14)	0.372(17)
	HRG	0.894(20)	0.885(18)	0.865(18)
	nSBM	0.9331(65)	0.9281(53)	**0.9137(64)**
	SPM	**0.9762(68)**	**0.941(11)**	0.862(18)
Fe	JC	0.5003(83)	0.5017(50)	0.5065(64)
	AA	0.5157(90)	0.5124(60)	0.5110(63)
	PA	0.8272(56)	0.8218(32)	0.8129(28)
	RA	0.5126(75)	0.5095(59)	0.5092(64)
	HRG	0.377(22)	0.380(12)	0.388(16)
	nSBM	**0.973**	**0.968**	**0.955**
	SPM	0.9602(42)	0.9485(34)	0.9218(48)

5 Outlook

The primary goal of this project was to study a physical system by the means of network science to find out which features of a network are related to notions and quantities in the physical world. For this purpose we represented atomic spectral data in terms of spectroscopic networks.

We showed in Chapter 2 that the basic topological properties like *network density*, *transitivity* and *degree assortativity* were not related to the physical properties of the atoms. Rather they seemed to express the way the data was obtained. This was especially apparent in the *degree distributions* of the different spectroscopic networks, which for networks of greater size and created from experimental had a form similar to the degree distribution of networks whose growth is guided by the *Preferential Attachment* building principle. We argued that the data taking process can be modelled in such a way, explaining why we did not see this form of distribution in networks based on theoretical data. Other than this similarity, we found no evidence that spectroscopic networks in general can be attributed to a class of networks that originate from generative models.

Furthermore, we studied in detail the differences between the structure of spectroscopic networks created from data that was experimentally obtained in contrast to such networks created from theoretical data. The theoretical data included not only electric dipole transitions, but also transitions of higher orders, which rendered the network based on such data much more regular. This is reflected in the higher *density* and *transitivity* and a lower *degree assortativity* (the latter only if we corrected for the bias introduced by the selection of energy states).

The results of structure analysis by the means of a community detection algorithm in Chapter 3 revealed that the partitioning of nodes, that the nested stochastic block model method found, relate to the quantum mechanic labels of the energy states in

terms of quantum numbers. This means that we were able to find the underlying symmetries of the physical system (*i.e.* quantum numbers relating to the symmetry groups of its Hamiltonian) solely by a data-driven network approach, without using the microscopic model of the atoms. Also, we showed that the networks, as its network structure is driven by the selection rules of electronic spectra, exhibits a *disassortative community structure*, *i.e.* the fraction of internal links inside a community is lower than the fraction of external links between different communities.

In Chapter 4 we demonstrated that we can not only recover known physical quantities by using this network approach, but we are also able to predict the existence of atomic transitions solely on the basis of the structural data of networks. We showed that several link prediction algorithms with different approaches were able to yield such predictions with high accuracy. The methods that were not assuming any organising principles performed more stable than those that did. The predictive power of the latter methods exhibited a robustness against an increasing dropout rate and the complexity of the considered atom. Considering both accuracy of the predictions and computational cost of the method, we argued that the structural perturbation method is most suited for link prediction in spectroscopic networks. Although their performance was not as consistent, the ability of methods that assumed underlying organising principles gave indications of the building mechanisms at work for this network type. The results for these methods support the findings of Chapter 3 that there is an disassortative hierarchical structure generated by the selection rules for atomic electron transitions. We argued that the advantages of this method for the prediction of transition over a prediction by the use of the selection rules is its generality with respect to the system and its ability to rank the results by a likelihood of existence.

With a short exception in Chapter 2 we used a representation of the spectral data by *simple graphs*, *i.e.* networks without node attributes, link weights, link directions or multiple links types. This representation made it possible to predict new transitions: since transitions were explicitly represented by links, we could use already existing link prediction methods to predict a physical entity. Thus, we see that the mapping from a system onto its network representation dictates what physical features can be studied. For further investigations of spectroscopic networks it would

therefore be a natural step to extend this representation, as more complex network models allow to store more information about the physical system. Or one could use a different representation to shift the focus of the investigation. A starting point could be the use of *weighted links*. One possibility for such weighting is to use transition probabilities (as proposed in Chapter 2) to be able to separate between the different transition types. This should for example lead to an enhanced predictive power of the link prediction methods. The most apparent obstacle to overcome when using transition probabilities as link weights are the enormous differences in value. A rescaling or a cut-off of the weights could overcome this problem, but this would constitute a heuristic ad hoc usage of domain knowledge, defying the generality of the data-driven approach. Introducing link weights also automatically leads to the question of whether we would be able to also predict these weights (either for existing links with unknown weights or together with the existence of unknown links). If there were feasible solutions for this weight prediction task, then another viable option for the choice of link weights may be the wavelengths of the respective transitions. This choice could possibly allow the prediction of the wavelength of an unknown transition. So far we were not able to study this in more detail, but there are existing methods for weight prediction, such as [40] or [47].

Besides predicting links and their weights we can use the structural data of the network to predict nodes, *i.e.* the existence of energy states. This is in general a hard problem, as in contrast to link prediction this means that we do not only predict values of entries of the adjacency matrix, but are extending the matrix and need to predict the values of the newly added columns and rows simultaneously. Such a prediction would be especially interesting in an extended network model such that we could not only infer the existence of nodes, but also node attributes, for example the energy of a previously unmeasured state. We started investigating this problem and applied the method proposed by Kim and Leskovec in [24] to spectroscopic networks. However, this method has several drawbacks, such as only allowing predictions of networks whose total number of nodes is a power of 2. Furthermore it could not recover nodes that we deleted for evaluation, so that we chose not to pursue the use of this method any further. My colleague D. Wellnitz developed a node prediction method based on the extension of the eigenvectors of the adjacency matrix. This methods yields better-than-random predictions accuracies under simplifying conditions. Detailed results of this method will be covered in his final thesis.

Spectroscopic networks have a disassortative community structure and we possess an microscopic model of the network structure. These are both properties that are not common in real-world networks and hence these networks can be used as a benchmark for the development of community detection methods that designed to find "anti-communities". See [28] for such a use of spectroscopic networks by S. Lackner *et al.*

Lastly, with respect to the community detection in spectroscopic networks we should note that we only presented the results of a single method in this particular representation. Applying other methods in this or other representation could lead to further insights about what rules the network structure is governed by, which in turn could lead to further insights into the connections between network structure and physical properties of atoms and their energy eigenstates.

Appendices

A Hierarchical Random Graphs

All of the following discussion and formulae are explicitly based on the work of Clauset *et al.* in [7] and the provided Supplementary Information. This appendix is intended as a summary of the features that are most relevant for link prediction in the framework of this method.

Definition of a Dendrogram In the hierarchical random graph framework proposed by Clauset *et al.*, the structure of a network is represented by a *dendrogram*. Consider a network graph $G = G(V, E)$ with n vertices. Then a dendrogram D is a binary tree with n leaves corresponding to the n vertices of G. Accordingly a dendrogram has $n-1$ internal nodes arranged hierarchically such that each internal node merges two branches of the level below it (or just two leaves in the lowest level). Thus each of these internal nodes corresponds - if one follows the hierarchy down to the lowest level - to a group of vertices of G that descend from it. In this model we will associate a probability p_r with each internal node r. This allows us to attribute a probability to every pair of vertices of G: given two vertices i, j of G, the probability p_{ij} that they are connected by an edge is $p_{ij} = p_r$ where r is their lowest common ancestor in the dendrogram D. The combination (D, p_r) of the dendrogram and the set of probabilites then defines a *hierarchical random graph* (HRG).

Likelihood of a Dendrogram There is no single unique representation of the network structure in terms of a HRG, but there are several possible for each network. Therefore we have to find a way to quantify how well a HRG is describing the network G and - in a second step - find the one that best suits the observed data. In this model the likelihood of the dendrogram D to describe the given network graph G is given by [7]

$$\mathcal{L}(D, \{p_r\}) = \prod_{r \in D} p_r^{E_r} (1 - p_r)^{L_r R_r - E_r}, \tag{A.1}$$

where E_r is the number of links in G whose endpoints (nodes) have the internal node r as their lowest common ancestor in D, and L_r, R_r the numbers of leaves in the left and right subtrees rooted at r.
For a fixed dendrogram this is maximised by the probabilities $\{\bar{p}_r\}$ given by

$$\bar{p}_r = \frac{E_r}{L_r R_r}, \tag{A.2}$$

which can be interpreted as the fraction of potential links between the nodes in the two subtrees that actually exist in G.
It follows that $\mathcal{L}$ evaluated at this maximum is

$$\mathcal{L}(D) = \prod_{r \in D} \left[\bar{p}_r^{\bar{p}_r} \left(1 - \bar{p}_r\right)^{1-\bar{p}_r}\right]^{L_r R_r}, \tag{A.3}$$

and its logarithm is

$$\log \mathcal{L}(D) = -\sum_{r \in D} L_r R_r h(\bar{p}_r), \tag{A.4}$$

where $h(p) = -p \log p - (1-p)\log(1-p)$ is the Gibbs-Shannon entropy function. From this entropy function stems the ability of this link prediction method to account for groups of nodes between which the connections are very likely or very rare. This entropy function maximises for the extreme cases of low and high probabilities, associated with disassortative or assortiative mixing behaviour.

Sampling dendrograms Clauset *et al.* state in that there are typically many dendrograms describing a network with equal likelihood. Therefore one should not only consider the best fit but take samples of the HRGs and average over their inherent link probabilities. Hence want to sample from all possible dendrograms with a probability according to their likelihood $\mathcal{L}$. For the sampling a *Markov chain Monte Carlo* (MCMC) method is proposed. The transition step from one dendrogram D to another dendrogram D' is done by a rearrangement of *subtrees*. The *subtrees* of an internal node r are the mutually independent parts of the dendrogram that are lower in hierarchy. Each internal node has three subtrees: two daughters and the subtree that descended form its sibling. These parts can be rearranged without changing the inner relationships. A graphical explanation of this is shown in the Supplementary Information of [7].

A step of of the Markov chain is then done as follows:

1. Choose an internal node of D uniformly at random
2. Choose one of the two alternate subtree configurations. Consider this new configuration the dendrogram D'.

This is the Markov part of the MCMC method. Following introduces the Monte Carlo characteristic: the new dendrogram is accepted as the new state in the Markov Chain according to the Metropolis-Hastings rule: accept if

$$\Delta \log \mathcal{L} = \log \mathcal{L}(D') - \log \mathcal{L}(D) \geq 0, \tag{A.5}$$

i.e. D' is at least as likely as D. Or, if this is not the case, accept with a probability

$$\exp\left(\log \Delta \mathcal{L}\right) = \frac{\mathcal{L}(D')}{\mathcal{L}(D)}. \tag{A.6}$$

This rule guarantees is combination with ergodicity of the steps in the Markov chain of this a probability distribution over dendrograms this is proportional to the likelihood, $P(D) \propto \mathcal{L}(D)$.

Link Probabilities To predict links, one samples dendrograms at regular intervals from the ones created by the Markov chain. For each sampled dendrogram D, one creates a Graph G and places a link between nodes i and j with probability $p_{ij} = \bar{p}_r$ with r being the lowest common ancestor of the nodes i and j in D. From these graphs one can compute averages of network statistics. Thus, for each non-existent link in G, *i.e.* for each unconnected pair of nodes i and j, one averages over the corresponding probabilities p_{ij} in each of the previously sampled dendrograms D to get the mean probability $\langle p_{ij} \rangle$ for the existence of this link. These non-existent pairs i, j can then be sorted in decreasing order of probability. This sorted list is the typical final reuslt of a link prediction algorithm.

Only the highest-ranked ones are thought to have a missing connection. Clauset *et al.* propose to use only the top 1 % of the predictions. We have not found a rigorous criterion for which fraction of predictions should be considered.

The implementation of this method was done with the programming code provided by Clauset *et al.* in combination with their publication [7].

B Structural Perturbation Method

The following discussion is entirely based on the work of Lü *et al.* in [29] and its intention is summarising the concepts of this method that are most relevant for this thesis.

For this method we will be considering the network in the framework of the adjacency matrix. Let E^T be the set of observed links, *i.e.* the training set, and let $A \in \{0,1\}^{N\times N}$ be the adjacency matrix of this observed network, where the element $A_{ij} = 1$ if the nodes i and j are connected by a link and $A_{ij} = 0$ otherwise. We then split the set E^T further into the two sets E_R and ΔE such that $E_R \cup \Delta E = E$ and $E_R \cap \Delta E = \emptyset$. Typically, the size of ΔE will be a tenth of E^T. A^R and ΔA are the corresponding adjacency matrices with $A = A^R + \Delta A$.

Next we express A^R in its eigendecomposition

$$A^R = \sum_{i=1}^{N} \lambda_i \mathbf{v}_i \mathbf{v}_i^\intercal \quad \in \{0,1\}^{N\times N}, \tag{B.1}$$

with λ_i and $\mathbf{v}_i$, $i \in \{1,\dots,N\}$ being the eigenvalues and eigenvectors of this matrix. At the heart of this method is that we can treat ΔA as a small perturbation to A^R. With this we can calculate the first order corrections $\Delta\lambda_i$ to the eigenvalues or A^R in its eigendecomposition (B.1)

$$\Delta\lambda_i \approx \frac{\mathbf{v}_i^\intercal \Delta A \mathbf{v}_i}{\mathbf{v}_i^\intercal \mathbf{v}_i}, \tag{B.2}$$

such that a perturbed matrix $\tilde{A}$ can be written as

$$\tilde{A} = \sum_{i=1}^{N} (\lambda_i + \Delta\lambda_i)\, \mathbf{v}_i \mathbf{v}_i^\intercal \quad \in \mathbb{R}^{N\times N}. \tag{B.3}$$

This can be considered as the linear approximation of the given network A for an expansion based on A^R and should be close to $A^R + \Delta A$, if the perturbation has not significantly changed the network structure.

Instead of having a similarity score for the likelihood of existence as it has been for the similarity-based link prediction methods, for the Structural Perturbation Method the non-observed links will be ranked by their corresponding entries in $\tilde{A}$. For the result not to be influenced by the randomness of the splitting of A^T into the two furhter sets A^R and ΔA, the algorithm will be run several times with differently sampled links and the ranked result will follow from the averaged adjacency matrix $\langle \tilde{A} \rangle$.

It should be noted, that in contrast to the other link prediction methods in this work, we can here not only vary the dropout ratio that determines the sizes of probe and training sets E^P and E^T, but also the the fraction of links in ΔE in the second splitting of E^T, which the authors refer to as p^H. We have not varied p^H in our experiments, but used $p^H = 0.1$ throughout the simulations, as proposed in [29].

To obtain the structural consistency σ_c we calculate the fraction of common links between the top-L ranked links E^L derived from $\tilde{A}$ and ΔE, where $L = |\Delta E|$

$$\sigma_c = \frac{|E^L \cap \Delta E|}{|\Delta E|}. \tag{B.4}$$

Bibliography

[1] K. M. Aggarwal and F. P. Keenan. 2006. Electron impact excitation of Fe XVI: radiative and excitation rates. *Astronomy & Astrophysics* 450, 3 (2006), 1249–1257. https://doi.org/10.1051/0004-6361:20054683 6

[2] A.-L. Barabási. 2016. Network Science. *Network Science, by Albert-László Barabási, Cambridge, UK: Cambridge University Press, 2016* (2016). 5, 13, 14, 15, 51

[3] A.-L. Barabási and R. Albert. 1999. Emergence of scaling in random networks. *Science (New York, N.Y.)* 286, 5439 (1999), 509–12. https://doi.org/10.1126/science.286.5439.509 13

[4] G. Bianconi and A.-L. Barabási. 2001. Bose-Einstein condensation in complex networks. *Physical review letters* 86, 24 (2001), 5632–5635. https://doi.org/10.1103/PhysRevLett.86.5632 1

[5] B. H. Bransden, C. J. Joachain, and T. J. Plivier. 2003. *Physics of atoms and molecules*. Pearson Education India. 11, 21, 30, 34

[6] A. D. Broido and A. Clauset. 2018. Scale-free networks are rare. (2018). arXiv:1801.03400 14, 15

[7] A. Clauset, C. Moore, and M. E.J. Newman. 2008. Hierarchical structure and the prediction of missing links in networks. *Nature* 453, 7191 (2008), 98–101. https://doi.org/10.1038/nature06830 56, 57, 76, 77, 78

[8] A. G. Császár and T. Furtenbacher. 2011. Spectroscopic networks. *Journal of Molecular Spectroscopy* 266, 2 (2011), 99–103. https://doi.org/10.1016/j.jms.2011.03.031 3, 15

[9] G. Drake. 2006. High precision calculations for helium. In *Springer Handbook of Atomic, Molecular, and Optical Physics*. Springer, 199–219. 17

[10] E. Dwek. 2016. Iron: A key element for understanding the origin and evolution of interstellar dust. *The Astrophysical Journal* 825, 2 (2016), 136. https://doi.org/10.3847/0004-637X 6

[11] P. Erdős and A. Rényi. 1960. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5, 1 (1960), 17–61. 13

[12] M. Faloutsos, P. Faloutsos, and C. Faloutsos. 1999. On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, Vol. 29. ACM, 251–262. https://doi.org/10.1145/316194.316229 14

[13] T. Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27, 8 (2006), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010 42, 43, 46

[14] S. Fortunato and D. Hric. 2016. Community detection in networks: A user guide. *Physics Reports* 659 (2016), 1–44. https://doi.org/10.1016/j.physrep.2016.09.002 28

[15] T. Furtenbacher, P. Arendás, G. Mellau, and A. G. Császár. 2014. Simple molecules as complex systems. *Scientific reports* 4 (2014), 4654. https://doi.org/10.1038/srep04654 3, 15

[16] T. Furtenbacher and A. G. Császár. 2012. The role of intensities in determining characteristics of spectroscopic networks. *Journal of Molecular Structure* 1009 (2012), 123–129. https://doi.org/10.1016/j.molstruc.2011.10.057 22

[17] J. Gao, D. Li, and S. Havlin. 2014. From a single network to a network of networks. *National Science Review* 1, 3 (2014), 346–356. https://doi.org/10.1093/nsr/nwu020 1

[18] R. Guimerà and M. Sales-Pardo. 2009. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences of the United States of America* 106, 52 (2009), 22073–22078. https://doi.org/10.1073/pnas.0908366106 40

[19] A. Halu, L. Ferretti, A. Vezzani, and G. Bianconi. 2012. Phase diagram of the Bose-Hubbard model on complex networks. *EPL (Europhysics Letters)* 99, 1 (2012), 18001. https://doi.org/10.1209/0295-5075/99/18001 1

[20] P. Holme and M. Huss. 2005. Role-similarity based functional prediction in networked systems: application to the yeast proteome. *Journal of the Royal Society, Interface / the Royal Society* 2, 4 (2005), 327–33. https://doi.org/10.1098/rsif.2005.0046 45

[21] L. Hubert and P. Arabie. 1985. Comparing partitions. *Journal of classification* 2, 1 (1985), 193–218. https://doi.org/10.1007/BF01908075 31

[22] O. Jitrik and C. F. Bunge. 2004. Transition probabilities for hydrogen-like atoms. *Journal of Physical and Chemical Reference Data* 33, 4 (2004), 1059–1070. https://doi.org/10.1063/1.1796671 6, 8, 9, 10, 19

[23] T. H. Johnson, S. R. Clark, and D. Jaksch. 2014. What is a quantum simulator? (2014), 1–13. https://doi.org/10.1186/epjqt10 39

[24] M. Kim and J. Leskovec. 2011. The Network Completion Problem: Inferring Missing Nodes and Edges in Networks. *SIAM International Conference on Data Mining* (2011), 47–58. https://doi.org/10.1137/1.9781611972818.5 73

[25] A. Kramida, Yu. Ralchenko, J. Reader, and NIST ASD Team. 2018. NIST Atomic Spectra Database (ver. 5.5.6), [Online]. Available: https://physics.nist.gov/asd. National Institute of Standards and Technology, Gaithersburg, MD. Accessed: 2018-04-07. 6, 9, 10, 19, 66

[26] A. E. Kramida. 2010. A critical compilation of experimental data on spectral lines and energy levels of hydrogen, deuterium, and tritium. *Atomic Data and Nuclear Data Tables* 96, 6 (2010), 586–644. https://doi.org/10.1016/j.adt.2010.05.001 19

[27] N. Kulvelis, M. Dolgushev, and O. Mülken. 2015. Universality at Breakdown of Quantum Transport on Complex Networks. *Physical Review Letters* 115, 12 (2015), 1–5. https://doi.org/10.1103/PhysRevLett.115.120602 1

[28] S. Lackner, A. Spitz, M. Weidemüller, and M. Gertz. 2018. Efficient Anti-community Detection in Complex Networks. *SSDBM'18: 30th International Conference on Scientific and Statistical Database Management.* (2018). https://doi.org/10.1145/3221269.3221289 To be Published. 74

[29] L. Lü, L. Pan, T. Zhou, Y.-C. Zhang, and H. E. Stanley. 2015. Toward link predictability of complex networks. *Proceedings of the National Academy of Sciences* 112, 8 (2015), 201424644. https://doi.org/10.1073/pnas.1424644112 62, 63, 64, 68, 79, 80

[30] L. Lu and T. Zhou. 2010. Link Prediction in Complex Networks: A Survey. *Physica A* 390, 6 (2010), 1150–1170. https://doi.org/10.1016/j.physa.2010.11.027 40, 43, 44, 45, 50, 51, 53, 57, 68, 69

[31] W. C. Martin and W. L. Wiese. 1996. Atomic, Molecular, and Optical Physics Handbook. *American Institute of Physics: Woodbury, New York* (1996). 29, 30

[32] M. E.J. Newman. 2002. Spread of epidemic disease on networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* 66, 1 (2002). https://doi.org/10.1103/PhysRevE.66.016128 13

[33] M. E.J. Newman. 2003. Mixing patterns in networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* 67, 2 (2003), 13. https://doi.org/10.1103/PhysRevE.67.026126 18

[34] M. E.J. Newman. 2003. The Structure and Function of Complex Networks. *SIAM Rev.* 45, 2 (2003), 167–256. https://doi.org/10.1137/S003614450342480 1

[35] M. E.J. Newman. 2010. *Networks: an introduction*. Oxford university press. 1, 5, 10, 11, 12, 13, 14, 17, 18, 27, 44, 45

[36] M. E.J. Newman and M. Girvan. 2004. Finding and evaluating community structure in networks. *Physical review E* 69, 2 (2004), 026113. https://doi.org/10.1103/PhysRevE.69.026113 28

[37] NIST ASD Team. 2018. NIST Atomic Spectra Database Help Page, [Online]. Available: https://physics.nist.gov/PhysRefData/ASD/Html/lineshelp.html. National Institute of Standards and Technology, Gaithersburg, MD. Accessed: 2018-05-13. 6

[38] T. P. Peixoto. 2014. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X* 4, 1 (2014), 1–18. https://doi.org/10.1103/PhysRevX.4.011047 28, 29, 37, 68

[39] T. P. Peixoto. 2017. Bayesian stochastic blockmodeling. *arXiv preprint arXiv:1705.10225;* (2017). To appear in "Advances in Network Clustering and Blockmodeling," edited by P. Doreian, V. Batagelj, A. Ferligoj, (Wiley, New York, 2018 [forthcoming]). 28

[40] T. P. Peixoto. 2018. Nonparametric weighted stochastic block models. *Physical Review E* 97, 1 (2018), 012306. https://doi.org/10.1103/PhysRevE.97.012306 73

[41] W. M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* 66, 336 (1971), 846–850. https://doi.org/10.1080/01621459.1971.10482356 30

[42] M. Rosvall, J.-C. Delvenne, M. T. Schaub, and R. Lambiotte. 2017. Different approaches to community detection. (2017). arXiv:1712.06468 27, 28

[43] M. Sales-Pardo, R. Guimera, A. A. Moreira, and L. A. Nunes Amaral. 2007. Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences* 104, 39 (2007), 15224–15229. https://doi.org/10.1073/pnas.0703740104 56

[44] M. A. Valdez, D. Jaschke, D. L. Vargas, and L. D. Carr. 2017. Quantifying Complexity in Quantum Phase Transitions via Mutual Information Complex Networks. *Phys. Rev. Lett.* 119 (Nov 2017), 225301. Issue 22. https://doi.org/10.1103/PhysRevLett.119.225301 1

[45] T. Vallès-Català, T. P. Peixoto, R. Guimerà, and M. Sales-Pardo. 2017. On the consistency between model selection and link prediction in networks. (2017), 1–12. arXiv:1705.07967 59

[46] P. Wang, B. Xu, Y. Wu, and X. Zhou. 2015. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences* 58, 1 (2015), 1–38. https://doi.org/10.1007/s11432-014-5237-y 52

[47] J. Zhao, L. Miao, J. Yang, H. Fang, Q.-M. Zhang, M. Nie, P. Holme, and T. Zhou. 2015. Prediction of links and weights in networks by reliable routes. *Scientific reports* 5 (2015), 12261. https://doi.org/doi:10.1038/srep12261 73

[48] T. Zhou, L. Lü, and Y.-C. Zhang. 2009. Predicting missing links via local information. *European Physical Journal B* 71, 4 (2009), 623–630. https://doi.org/10.1140/epjb/e2009-00335-8 53, 54, 55

Affidavit

I affirm that this thesis was written by myself without any unauthorised third-party support. All used references and resources are clearly indicated.

Ich versichere, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, May 30, 2018 .....