Department of Physics and Astronomy

University of Heidelberg

Master thesis

in Physics

submitted by

Wellnitz David

born in Rheda-Wiedenbrück

2018

A Network Approach

to

Atomic Spectra

This Master thesis has been carried out by Wellnitz David

at the

Physics Institute

under the supervision of

Herrn Prof. Weidemüller Matthias

Ein netzwerktheoretischer Ansatz zur Untersuchung von Atomspektren:

In dieser Arbeit untersuchen wir Atomspektren mithilfe von Konzepten der Netzwerktheorie. Wir erzeugen ein sogenanntes *spektroskopisches Netzwerk* aus Spektraldaten, indem wir Zustände als Knoten und optische Übergänge als Kanten definieren. Wir zeigen, dass spektroskopische Netzwerke aufgrund der Parität in guter Näherung bipartit sind. Mithilfe der Methode der *Community Detection* finden wir Gruppen von Knoten, welchen identische Quantenzahlen zugeordnet werden können. In Thorium II zeigen sich weitere Hierarchieebenen, die sich als Cluster in den entsprechenden Energien wiederfinden. Innerhalb dieser Cluster haben die Zustände eine ähnliche Elektronenkonfiguration. Aus diesen Resultaten entwickeln wir eine Methode, um Eigenschaften einzelner Zustände vorherzusagen. Für Thorium II bestimmen wir so die richtige Parität in 100 % und das richtige J in 95 % der Fälle.
Die Methode der *Link Prediction* wird angewandt, um Übergänge vorherzusagen. Wir testen die *Structural Perturbation Method* und das *Nested Stochastic Block Model* und werten sie durch zufälliges Entfernen von Kanten aus. Mithilfe der Structural Perturbation Method können je nach Atom 20 % bis 40 % der Übergänge mit geringer Fehlerquote wiedergefunden werden. Wir entwickeln eine neue Methode, um die Übergänge zu bislang nicht bekannten Zuständen anhand ihrer erwarteten quantenmechanischen Eigenschaften vorherzusagen.

A Network Approach to Atomic Spectra:

In this thesis we investigate atomic spectra using network theory. The spectroscopic data of an atom is mapped onto a network by identifying nodes with energy levels and links with optical transitions. These so-called spectroscopic networks are almost bipartite, which is ascribed to the parity symmetry. We apply community detection to these networks and show that the communities found correspond to known quantum numbers. For thorium II we demonstrate that the states of additional communities detected form one or two clusters in the energy domain. These clusters correlate with the dominant configurations of the states. We test the ability to use the network to predict quantum mechanical properties of individual states and show that for thorium II parity and J can be predicted with 100 % and 95 % accuracy respectively.
We show that new transitions can be predicted using state of the art methods of link prediction. We benchmark this by a random dropout and demonstrate that depending on the atom 20 % to 40 % of the transitions can be recovered with few errors using the structural perturbation method. We develop a method that can predict the transitions to new states, for which structural information is given.

Contents

1 Introduction

In this thesis we will discuss a novel, network based approach to analyze atomic spectra. This approach will shed a new light on some important questions in spectroscopic analysis and show the usefulness of networks as a tool for analyzing such systems.

Atomic spectra are of great interest in many areas of physics and chemistry, as they have been established as a unique fingerprint of atoms and as a way to analyze the atomic structure (see, e.g., Cowan [1981]). One example for current interest in spectra is the analysis of thorium, which Peik and Tamm [2003] proposed for the development of a precise nuclear clock. In order to excite the nuclear transition, Herrera-Sancho et al. [2013] proposed to use a coupling of this transition to electronic states. To find such a coupling, detailed knowledge about the spectrum of thorium is needed, and Redman et al. [2014] measured many transitions of thorium. Although a large amount of data is available, it is impossible for current theory to relate these spectra to the underlying microscopic physical laws, and even a simple question like "What are the angular momenta characterizing the measured states in thorium II?" is impossible to answer (see Redman et al. [2014]).

The information that can be extracted from the spectral measurement is hence limited to the transitions and the energy levels. These can be inferred by matching energy differences in the transitions to energy differences of known energy levels (see Redman et al. [2014]). For the transitions intensities are known in addition to the wavelengths. Although by the Ritz principle all pairs of states are connected by transitions (see, e.g., Cowan [1981]), the intensities of these transitions can differ by many orders of magnitude. This means that in practice only few transitions are relevant for each state. These transitions are usually electric dipole transitions, for which well known selection rules can be given, as explained, e.g., in the textbook by Bransden et al. [2003]. The spectroscopic data is generally given by a set of energy levels and the corresponding transitions connecting these.

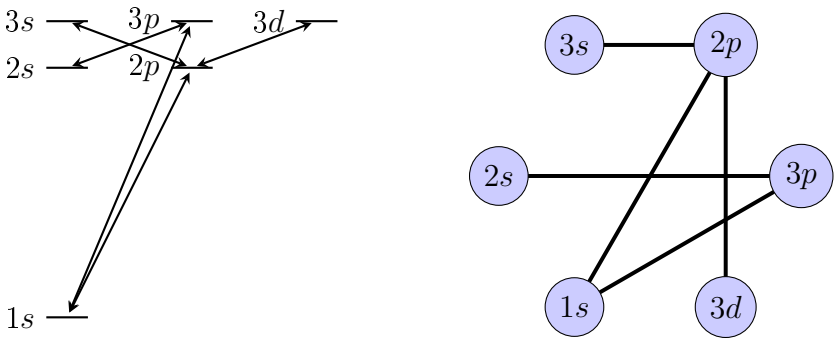
This type of data and the representation as a Grotrian diagram in particular (see figure 1.1) can be represented by a network in an intuitive manner. Networks consist

of a set of nodes (corresponding to states) and connections between them called links (corresponding to transitions), as described for example by Newman [2003]. This straightforward identification was proposed by Császár et al. [2016].

Over the last decades, network theory has evolved into a powerful tool to analyze complex systems. The information about the underlying system is represented by the connectivity pattern of the links of a network (see Newman [2003]). Various features of these connectivity patterns have been analyzed and related to underlying principles of the systems they describe (see e.g., the textbooks by Newman [2010] and Barabási et al. [2016]). Prominent examples can be found in different fields, such as cell biology, neuroscience, sociology etc.

One question that has sparked interest in various fields is: "What are groups of similar nodes?". Girvan and Newman [2002] proposed a method of community detection to answer this question. Since then, a great manifold of algorithms have been developed to tackle this problem. These have been reviewed by Fortunato [2010], who also gave an introduction into the topic Fortunato and Hric [2016]. Traditionally these approaches search for assortative communities, but there are also frameworks for general communities given by the stochastic block model, which was described by Peixoto [2017] and Abbe [2017].

Another question, which has been analyzed in great detail, is: "How can the data be extrapolated to predict new links?". Lü and Zhou [2011] provided a review on basic approaches of this problem. Sophisticated algorithms have been developed to tackle this problem, for example by Lü et al. [2015] and Guimerà and Sales-Pardo [2009]. This question has been extended to the prediction of links for new nodes. Lika et al. [2014] reviewed algorithms specifically for the recommendation of new



(a) Grotian diagram (b) Corresponding network

Figure 1.1: Correspondence between Grotian diagram (left) and network (right).

products and Hric et al. [2016] introduced methods based on the stochastic block model.

Despite all this analysis and the many successful applications, research applying network theory to physical systems is sparse. Halu et al. [2012] analyze the phase transition of the Bose-Hubbard model, where the interaction is given by a network instead of the lattice; Kulvelis et al. [2015] analyze the efficiency of quantum transport on tree like networks. As most of the research done at the intersection of physics and networks, these approaches do not try to gain new insights into physics based on networks, but merely use networks as the underlying topology of the physical system.

We only know of two approaches to gain new insights into a system based on network science. Valdez et al. [2017] used network theory to analyze the quantum phase transition of the one-dimensional Bose-Hubbard model. They used the particles as the nodes of their network and the mutual information as weighted links. They found that several, typical network properties show a characteristic behavior at the quantum phase transition of the system.

The other approach is an analysis of molecular spectra by Császár et al. [2016]. They made some basic network considerations in order to gain structural insights into molecular spectra, but only weak conclusions were drawn. Instead they focused on using networks as a tool to find a correct mapping between experimental and theoretical spectroscopic data.

In this thesis we aim to do an elaborate analysis of the networks generated from atomic spectra. We will use the methods of community detection and link prediction to tackle the following questions about atomic structure and atomic spectra:

- What are groups of states with similar properties?
- How do these groups relate to quantum mechanics?
- Which additional transitions are possible, and which new states could they include?

This analysis will go far beyond the analysis done by Császár et al. [2016], who considered the degree distribution, centralities and the modularity (see Clauset et al. [2004]) for community detection. We will discuss in section 2.3 that this community detection is not appropriate for spectroscopic networks due to their disassortative structure.

In order to benchmark the different methods, we will analyze the spectra of hydrogen, helium, iron and thorium. Hydrogen and helium are the simplest two atoms and are well analyzed theoretically. Thus they are perfect systems for benchmarking the methods and prevent influence from experimental bias. Iron and thorium II are complementary to that: Their atomic structure is way more complex, but much experimental data is available. Therefore these atoms are also the atoms, for which network theory can potentially offer additional insights.

The thesis is organized as follows: First, we analyze the structure of the atoms in chapter 2, then we predict new transitions and states in chapter 3. We will start by introducing a natural mapping between spectra and networks in section 2.1. In section 2.2, we will discuss how to use the intensities and network structure to find the parity symmetry. In order to find further structure in the network, community detection will be introduced in section 2.3. For helium, iron and thorium the communities found with this algorithm will be compared to the known quantum numbers and a correlation is found. For iron and thorium additional structure is found and analyzed using the energies of the states in section 2.4. To close the first chapter, we will show how the results can be used to predict quantum numbers of individual states in section 2.5.

In the second chapter we will use the network to make predictions about the spectrum. We will start in section 3.1 by employing state of the art methods for link prediction in order to predict transitions between two known states. To conclude, in section 3.2 new methods that can be used to predict new nodes in a network will be proposed, and these methods will be tested by predicting new states for the atom.

For the prediction of new states, the methods developed were also tested against various other real world networks and synthetic networks. With exception of the Zachary karate club (see Zachary [1977]), which has only 34 nodes, the method performed well on all networks tested. Furthermore the algorithm can be straightforwardly generalized to weighted networks. Testing this generalization has lead to inconclusive results, as it seems to work well for hydrogen and iron, but poorly for helium, carbon and the neural network of C. elegans.

The communities that are found and lead to a finer splitting than the quantum numbers parity and J were also analyzed for iron. Here, for some communities a relation between energies and the term symbol was found, which could not confirmed for thorium II. In addition, the energies show a less clear structure, sometimes

splitting into many different clusters. In addition tests comparing the groups to the entire spectrum were made. The latter two steps are shown compactly in the appendix A.

2 From Network Properties to Quantum Mechanics

The goal of this chapter is to employ network theory in order to answer the first two questions posed in the introduction: "What are groups of states with similar properties?" and "How do these groups relate to quantum mechanics?". We will find groups of similar states – called communities – in the network and show that their quantum mechanical properties are also similar.

In order to quantify similar quantum mechanical properties, we will use quantum numbers. Quantum numbers are operators with discrete eigenvalues that commute with the Hamiltonian. For hydrogen and helium, these are the parity, the orbital angular momentum L, the spin S and the total angular momentum J; for iron and thorium only parity and total angular momentum are known quantum numbers. To describe additional structure that is not captured by known quantum numbers, observables that do not commute with the Hamiltonian could be used as well, such as the electron configuration for thorium.

We will show that the communities are strongly correlated with known quantum numbers and that for thorium II, additional structure found by the communities is associated with the energies of the states. It will further be possible to predict the values of the known quantum numbers for individual states with high accuracy using the network structure and the quantum numbers of other states. For example, not every state of thorium II has a known value J, thus a prediction can be used to assign these values.

First, a way to map spectra onto networks will be introduced in section 2.1. Next, the structure of this network together with the intensities will be employed to assign a parity to each state in section 2.2. In section 2.3, community detection will be used in order to find the communities of similar states mentioned above. These communities will be set into relation with the different quantum numbers, and for helium we will find that there is a simple correspondence between communities and quantum numbers. For iron and thorium additional communities on top of the

quantum numbers are found. This additional structure is analyzed in section 2.4 with respect to the energies of the different states, where a relation between energy and community is found. These results will be used in section 2.5 to predict the quantum numbers for individual states.

2.1 Generating Networks from Spectra

In this section, some basics that are needed in the remainder of the thesis will be discussed. We will start by discussing the relation of networks to linear algebra. Then, we propose a mapping from spectral data to a network, and show how additional information can be encoded in weights. In the end, we give an overview over the spectroscopic data used to create the networks.

2.1.1 Networks and Linear Algebra

A general introduction into network theory can be found for example in the book by **?**. We will focus on the relation to linear algebra and start by introducing a mapping from networks to matrices. This relation to linear algebra enables us to transfer well known concepts like eigenvalues and eigenvectors onto networks and then atomic spectra. It will further enable a more general understanding of weights.

A simple network consisting only of nodes and links can be translated into the so called adjacency matrix. The adjacency matrix $\mathbf{A}$ is a $n \times n$ symmetric matrix, where n is the number of nodes in the network. The elements a_{ij} are given by:

$$a_{ij} = \begin{cases} 1 & \text{if there is a link connecting nodes i and j} \\ 0 & \text{else} \end{cases} \tag{2.1}$$

Each line and column of this matrix correspond to one node in the network. The eigenvectors and eigenvalues of this matrix capture many important properties and their analysis is its own field called spectral graph theory. Literature on this field can be found for example by Chung and Graham [1997] and Cvetković et al. [1997]. The eigenvectors will play an important role for the prediction of quantum numbers, links and nodes in the sections 2.5, 3.1 and 3.2.

To include further information into the network one can go from a simple network to a weighted network by assigning positive real numbers – so called weights – to the links of the network. This can be done by a function $w(i, j)$. The adjacency

matrix then generalizes to the weight matrix $\mathbf{W}$:

$$w_{ij} = \begin{cases} w(i,j) & \text{if there is a link connecting nodes i and j} \\ 0 & \text{else} \end{cases} \tag{2.2}$$

From here, one can see that a zero weight should correspond to a non-existent link, and low weights should correspond to less important links.

2.1.2 Network Creation

The natural approach to create a network from a spectrum was proposed by Császár et al. [2016] for molecules and is sketched in figure 1.1. In its essence it is the Grotian diagram without energies and directions. The nodes of the network are the energy eigenstates of the atom, and two nodes are connected by a link, if there is an optical transition between the respective states. The network generated this way is called *spectroscopic network*.

In order to encode further structure, a weighted network can be considered. As discussed above such a network could be generated from a (positive valued, symmetric) characteristic matrix of the atom with rows and columns corresponding to states. This matrix should be easy to access experimentally and the weights should vanish for non-existent transitions. The necessary physical background can be found in Bransden et al. [2003].

Császár et al. [2016] choose the transition intensities to generate such a matrix. As these do not directly capture the structure of the atom we propose a different matrix: the dipole matrix $\vec{\mathbf{D}}$. This matrix is directly related to the distance vector $\vec{r}$ between the two states: $\vec{\mathbf{D}} = e\vec{\mathbf{r}}$, and hence captures important information about the structure.

In order to calculate the dipole matrix element, it needs to be related to observables in the spectrum. The emission intensity is given by the Einstein coefficient A_{ik}. In the dipole approximation these are given by:

$$A_{ik}^{E1} = \frac{\omega_{ik}^3}{\pi\varepsilon_0 c^3\hbar}\, {}_i\langle\vec{\epsilon}\vec{D}\rangle_k \tag{2.3}$$

$\vec{\varepsilon}$ is the polarization of the light. A detailed derivation of these formula can be found in Bransden et al. [2003].

Thus the dipole matrix element can be related to the observed intensity A_{ik} and

wavelength λ_{ik} as $D_{ik} \propto A_{ik}^{E1} \cdot \lambda_{ik}^3$. And thus we can use $A_{ik} \cdot \lambda_{ik}^3$ as a weight. The additional factor λ_{ik}^3 is caused by the vacuum coupling of the transition, which is stronger for short wavelengths. As λ is finite for all relevant cases, the dipole matrix element has the desired property to vanish for hard to measure transitions. The weight choice $A_{ik} \cdot \lambda_{ik}^3$ has the added benefit that also non-dipole transitions get a small weight assigned.

2.1.3 Data

At this point a short summary of the data is given. This will be helpful to interpret the results in the following sections, as some features of the structure found in the network will be traced back to the particular choice of data.

For hydrogen, a purely theoretical dataset will be used, which was calculated by Jitrik and Bunge [2004]. This dataset contains 289 different energy levels and around 15 800 transitions. As hydrogen is well known, this dataset allows us to compare our methods against nature directly. All electric and magnetic dipole, quadrupole and octupole transitions were calculated, but only electric dipole transitions are used for the network, because these transitions are the strongest and would most likely be found in a measurement. In the choice of states there is a discrepancy between high and low angular momenta: For $l < 4$ more than 20 states were calculated for each combination of l and j, for $l \geq 4$ only 3 states were calculated for each combination of l and j. This discrepancy will lead to some artifacts in the further analysis.

The data for the spectra of iron, thorium and helium is taken from the NIST database Kramida et al. [2018]. For iron and thorium II the transitions were found only by experimental methods. Using the Ritz method energy levels were calculated from these spectra and used to determine the transition wavelengths more accurately. With this method 846 levels and around 10 000 transitions were found for iron and 516 levels with around 6 500 transitions were found for thorium. All of the iron-states and many of the thorium-states were also fitted to theoretically calculated energy levels, thus assigning further quantum numbers. This assignment of quantum numbers beyond parity and J should be seen as a good guess and not as a physical truth.

The helium spectrum is composed of both calculated and measured data, overall with 191 levels and 2 300 transitions. The measured data is also used for Ritz calculations and fitted to theoretical values. For states with $2 \leq L \leq 6$ and $n \leq 10$ ($n \leq 9$ for $L = 7$) all electrical dipole transitions including singlet-triplet transitions

have been calculated. Since most of the data is calculated, the helium network offers the possibility to benchmark the methods on a system with no experimental data selection bias. It will just be impossible to distinguish between singlet and triplet for large L.

All in all we thus have theoretical data of simple atoms as benchmark systems and experimental data for complex atoms. Both sets contain a bias by the choice of data: The theoretical datasets introduce artificial cutoffs on which states and transitions should be considered, the experimental data is biased towards certain wavelengths and states close to the ground state.

2.2 Weights and Bipartivity – Identifying Parity from Network Structure

In this section the weights are explored further and the parity is compared to a network property called bipartivity. A bipartite network can be split into two types of nodes A and B and each link connects a node of type A to a node of type B. This will be identified with the parity, as each transition changes the parity of the state.

This is a proof of principle that spectroscopic networks are a good representation of the data and can lead to an intuitive understanding of certain quantum mechanical properties. Further we will prove that the method proposed is successful under reasonable conditions, so that it offers a formalized tool to uncover the parity.

In the first part, the weights will be analyzed for the three networks of hydrogen, helium and iron with respect to the different transition types. In the second part the algorithm to find parity is described and it is proven that this algorithm is successful under reasonable conditions.

2.2.1 Weight Analysis

In order to analyze the weights a histogram of the weights in the network is drawn. The results can be seen in the figures 2.1, 2.2 and 2.3 for hydrogen, helium and iron respectively. For thorium only dipole transitions are measured, so thorium will not be considered in the following analysis.

For all three atoms we observe a hierarchy between the different types of transitions: Electric dipole transitions (E1) are the ones with the highest weight, followed by electric quadrupole (E2) and magnetic dipole (M1), both of which have about equal weights. Electric octupole (E3) and magnetic quadrupole (M2) transitions are weaker, and the weakest transitions considered were magnetic octupole (M3) transitions. This structure can be seen for all three atoms, but it is strongest for atoms with higher nuclear charge like iron. This could also be the reason why for thorium, which has a very high nuclear charge, only E1 transitions have been observed.

In figure 2.1 an overlap can be seen between the peaks belonging to the different transition types. For some E1 transitions $A_{ik} * \lambda_{ik}^3$ is as low as $1 \times 10^9\,\mathrm{nm^3 s^{-1}}$, whereas we can find E2 transitions with $A_{ik} * \lambda_{ik}^3$ up to $1 \times 10^{11}\,\mathrm{nm^3 s^{-1}}$. For helium this overlap is a lot smaller and for iron there is almost no overlap. For these two

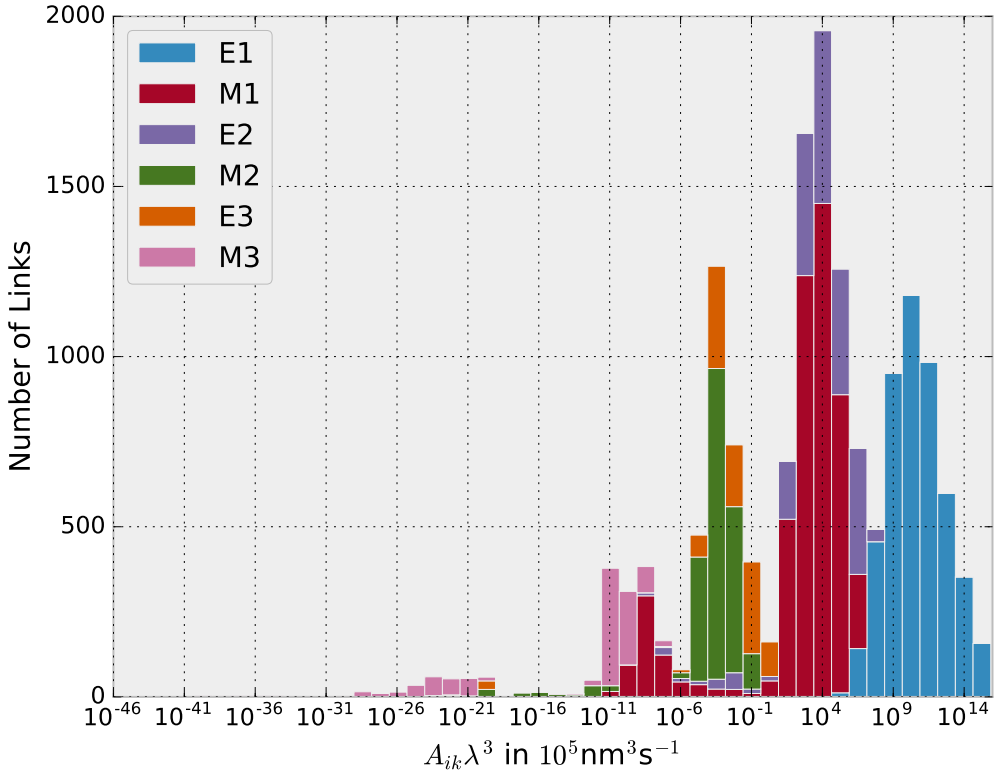


Figure 2.1: Weights of the hydrogen transitions by transition type. Here E and M are electric and magnetic; 1, 2 and 3 are dipole, quadrupole and octupole transitions respectively.

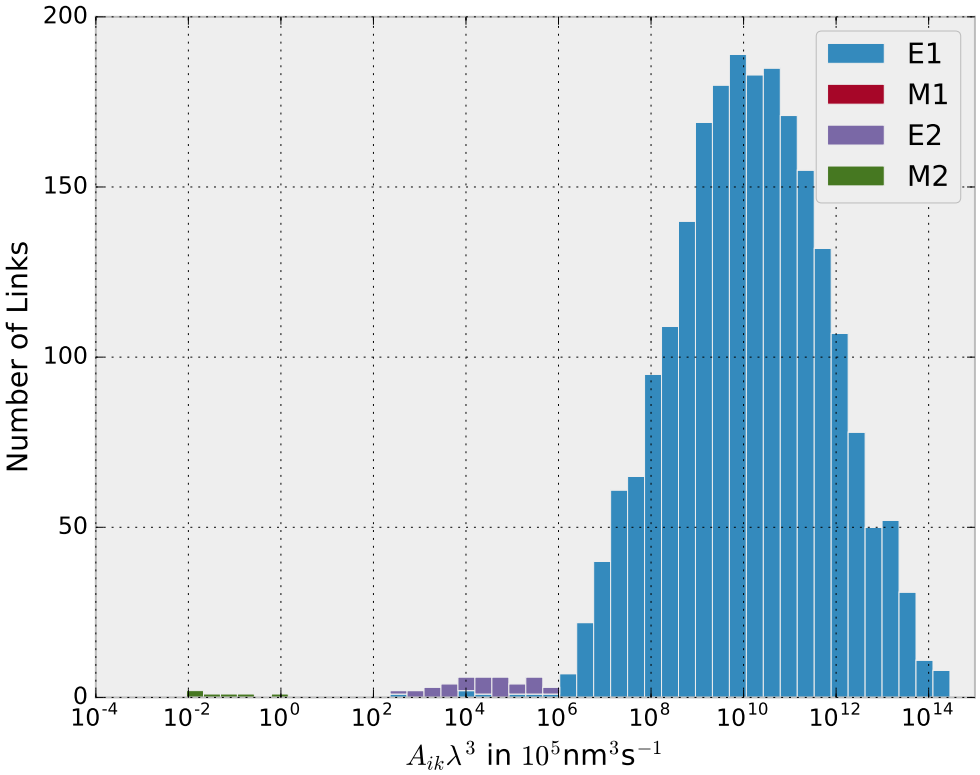


Figure 2.2: Weights of the helium transitions by the transition type.

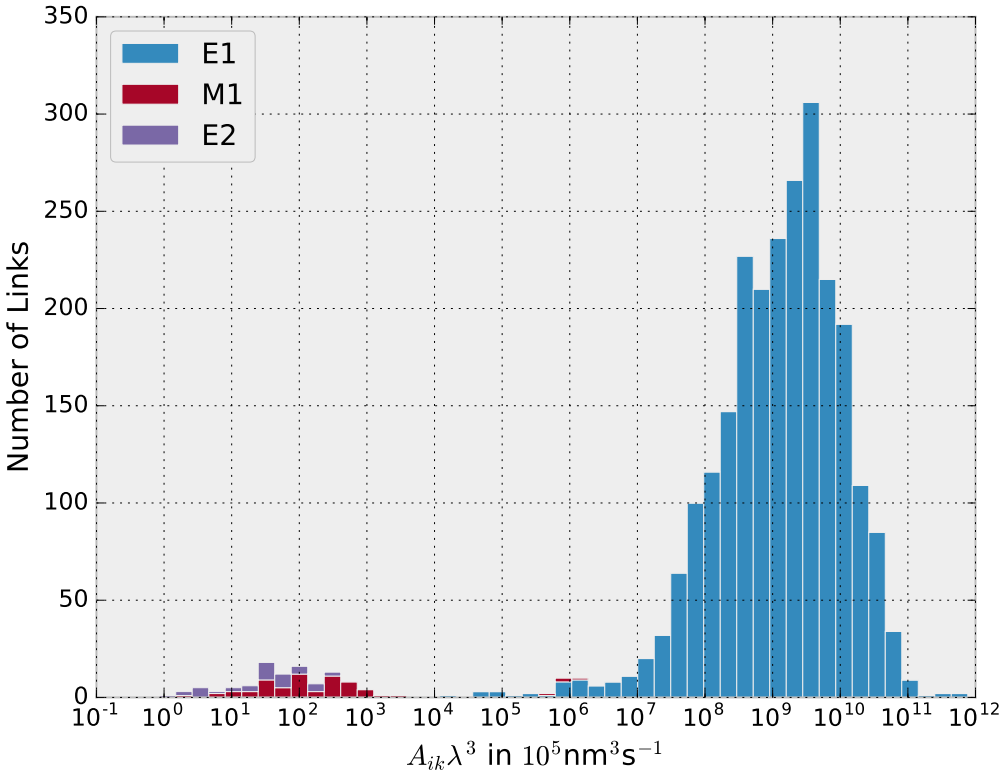
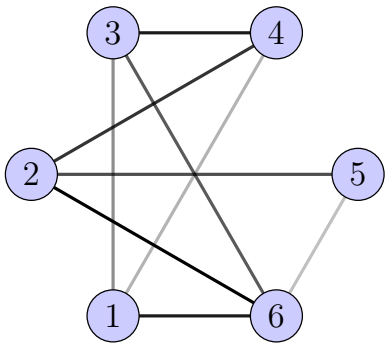
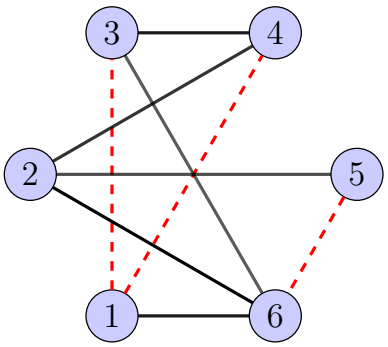


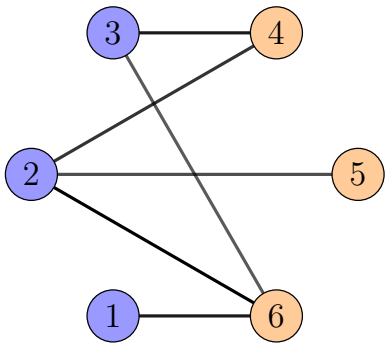
Figure 2.3: Weights of the iron transitions by transition type.



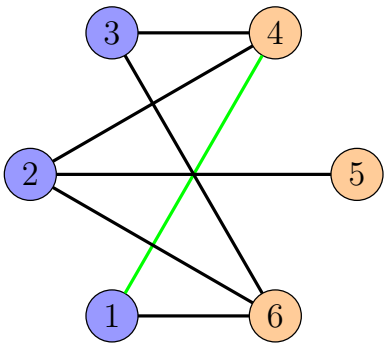
(a) The network given is not bipartite.



(b) We cut links by ascending weight until the network is bipartite.



(c) The new network is split into two bipartite groups.



(d) All lines connecting the two groups are added again.

Figure 2.4: Sketch of the algorithm to identify the dipole transitions.

atoms not all transitions have been measured, so it might well be that the overlap of these peaks would be larger, if all transitions were considered.

Thus we found a hierarchy in the different transition types: $E1 > (E2, M1) > (E3, M2) > M3$, but there is some overlap between the peaks for the different transition types.

2.2.2 Finding Parity

In this section we propose a method to distinguish the two types of nodes induced by the parity. For this we will need the concept of bipartivity introduced above. On top of that we will use that for a connected, bipartite network the splitting of the network into the two subsets A and B is unique.

We start with some observations. The selection rules enforce that each transition

of the types E1, M2 or E3 connects two links of different parity. Therefore, if only these transitions are considered, the network is bipartite and the subsets A and B split the network by parity. This was also found by Császár et al. [2016] for molecules, where only E1 transitions are measured and thus the network is always bipartite. Furthermore each M1, E2 or M3 transition connects two states of equal parity. Thus, if the network is connected, each of these transitions will destroy the bipartivity: Without the additional link the choice of A and B is unique, but after adding this link there is an internal link in one of the groups.

The task of finding parity is thus related to finding a connected, bipartite network that contains only E1, M2 and E3 transitions. This is achieved by the algorithm described in figure 2.4:

1. Is the network bipartite?
   - Yes: Go to step 4!
   - No: Sort the links by weight and continue! (2.4a)
2. Find the link with the lowest weight which can be removed without disconnecting the network and remove it. (2.4b)
3. Is the network bipartite?
   - Yes: Continue!
   - No: Go back to step 2!
4. Find the bipartite subsets A and B! (2.4c)
5. Add all links connecting A and B! (2.4d)

The subsets A and B now correspond to even and odd parity. To identify which one is even and which one is odd, one state has to have a known parity, which is usually given for the ground state. Furthermore, all parity changing transitions (E1, M2 or E3) are in the network. Since M2 and E3 transitions are significantly weaker than E1 transitions (see figure 2.1), the E1 transitions can be identified by finding separate peaks in the weight histogram.

We now proof that this algorithm works, if for any partition of the nodes into two sets the strongest link connecting these two sets is an E1, M2 or E3 transition, the two bipartite subsets A and B will correspond even and odd parity.

Proof: By construction the strongest links connecting two groups are never removed. These transitions change the parity by condition and connect the network by definition. If all but these links are removed the bipartite sets are obviously given by parity. Additional links cannot change this, as they will either lead to the same bipartite sets or be deleted. Thus the bipartite sets correspond to parity. In step 5 all links that change parity are added back, so that all E1, M2 and E3 links are part of the network.

q.e.d.

The condition that the strongest link connecting any partition of the nodes is an E1 transition is reasonable from a physics perspective. As shown in figure 2.1 even for hydrogen there are several thousand E1 transitions stronger than any other transition. These transitions should span the entire network and thus any partition of nodes is connected by one of these transitions. The only situation in which this algorithm should fail is if for a particular node due to experimental interest no E1 transition was measured. The influence of such an error would be limited to this particular node, so this is a minor consideration.

2.2.3 Conclusion

We have found that there is a clear hierarchy in the weights between the different transition types:

$$E1 > (E2, M1) > (E3, M2) > M3 \tag{2.4}$$

This hierarchy is broken by an overlap of the respective peaks in the weight histograms.

The parity of the different states can further be identified with bipartite subsets of the respective network, if weak links are not considered. This way not only the two parity subsets, but also which transitions belong to the types E1, M2 and E3 can be identified.

Even though this task can also be performed without network considerations, the approach presented here offers a new perspective on bipartivity. Furthermore it shows how simple network properties can be related to quantum numbers and how to use this relation. It shows that the network picture of the spectrum is reasonable and can lead to physical insights as well as some intuition about spectral properties.

2.3 Community Detection – Network Structure meets Quantum Mechanics

In the last section we were able to identify the parity of each node and the parity changing transitions using the structure of the network together with weights. In this section we ask: "Can we find further groups with similar properties and how are they related to quantum mechanics?" In order to answer this question we will use community detection. We will show that communities are groups of states with similar quantum numbers.

First the nested stochastic block model is introduced as a method to find communities in a network. These communities will be analyzed for the atoms helium, iron and thorium first qualitatively, then they will be quantitatively compared to the different quantum numbers. From this section onwards the weights will not be considered anymore, as they complicate the analysis and good results are achieved without them.

2.3.1 The Stochastic Block Model

Most community detection algorithms look for communities which have a lot of links within the communities (assortative structure). This strategy is not promising here, as spectroscopic networks are almost bipartite and assortative communities would have to mix the two bipartite group and thus find groups with mixed parities. This makes little sense from a quantum mechanical point of view. Therefore, a community detection algorithm should be used, which can find disassortative structure in the network.

This problem can be solved by using the stochastic block model (SBM), which is described in depth by Peixoto [2017]. The underlying assumption of the model is that there are different types of nodes. These types of nodes can be grouped together, and whether or not there is a link between two nodes is determined by the type of the two nodes. Here is an example: Consider three nodes l, m and n with one link connecting l and n and another one connecting m and n. The types could now be assigned as: $l, m \to A$ and $n \to B$ with the rule that all nodes of type A connect to all nodes of type B and no other connections are in the network. Therefore the SBM also seems like a reasonable approach to identify quantum numbers, as these can be seen as types of nodes and their connectivity is described by the selection

rules.

The goal of the SBM is to identify the underlying types of the nodes in a way that not each node gets its individual type, but still in a way that allows the reconstruction of the network from the types of the nodes. Thus we need to decide what structure is caused by noise (for example because not all links were observed), and what structure is caused by the underlying principles of the network. This decision can be made using Bayesian statistics with non-informative priors. A set of communities is a good model for the system if the amount of bits needed to describe the system with this model – the description length – is short. A model can shorten the description length, as many possible link combinations are ruled out by the model. Thus, in order to find the best set of communities, the description length needs to be minimized. The exact statistical foundation is explained in detail in Peixoto [2017].

The cost of the statistical description is a high computational cost, but as only small networks of 100 to 1 000 nodes are considered this does not matter. Furthermore, the SBM has problems finding small communities as they are misinterpreted as noise. This problem can be solved by introducing a hierarchical community structure with the nested stochastic block model (NSBM). This hierarchy can be imagined at the example of spectroscopic networks: On the lowest level, the network is split into two different groups according to the node parity. In a lower level, the network is then also split according to J into communities that now have same parity and same J. In a last step the communities are now split by L as well, leading to communities with same parity, J and L on the lowest hierarchy level. This insertion of hierarchical structure lowers the minimum size of a group from $\mathcal{O}(\sqrt{N})$ to $\mathcal{O}(\ln N)$ for N nodes in the network.

Even though weighted versions of the SBM were developed by Clauset et al. [2013] and Peixoto [2018a], implementing these would be beyond the scope of this thesis. Furthermore, even the unweighted version leads to good results.

The following specifications are used in the analysis: In order to find the best hierarchical structure, the software graph-tool by Peixoto [2018b] was used. The function `minimize_nested_blockmodel_dl` uses a Markov chain Monte Carlo algorithm to find the model with the minimal description length. As this algorithm is stochastic, the function was run 100 times and the result with the shortest description length was used in the further analysis. For helium the algorithm found the same minimum many times, indicating this is an absolute minimum of the data. For the other atoms, each minimum has only been found once, so even after 100 runs

we probably did not find the absolute minimum.

In the section "To sample or to optimize" in Peixoto [2017] it is discussed whether the best possible community model should be used or all models should be weighted with their respective likelihood. As in this chapter one specific community structure should be chosen for the network, only the most likely model will be chosen for this. In later sections, when predictions are made, different models will be sampled instead with their respective likelihood.

All in all the SBM is a community detection method, which can detect assortative and disassortative structures. In order to improve the resolution the nested version will be used, which also finds an additional hierarchy.

2.3.2 Results

The results of this approach are shown in figures 2.5, 2.6 and 2.7. In each graph, the community structure is encoded in the spacial positions of the nodes and the quantum mechanical structure is encoded in the style of the nodes. For helium the position only encodes the lowest hierarchy level with the smallest communities, whereas for thorium and iron the different levels are shown: The highest level – which corresponds to the parity splitting – is indicated by the left-right separation, the lowest level are the very close groups of nodes and the (for iron two and for thorium one) intermediate levels are given by groups that are closer together. Further they have been arranged in a way that from bottom to top the total angular momentum J is in general increasing. For helium the entire term symbol is given in the graph, whereas for iron and thorium only the quantum numbers parity and J are given. These are the only known, non-trivial observables which commute with the Hamiltonian and thus the only known good quantum numbers.

Helium

In figure 2.5 can be seen that the spacial groups correlate almost perfectly with the colors and the symbols of the nodes. This means that in general there is a strong correspondence between the community and the quantum numbers orbital angular momentum L and total angular momentum J. For small angular momentum $L \leq 2$ the communities also correlate with the shape of the node. This means S is found as well. For higher angular momentum $L > 2$ the total spin is not identified by the community detection algorithm anymore. This can be seen as the singlet and the

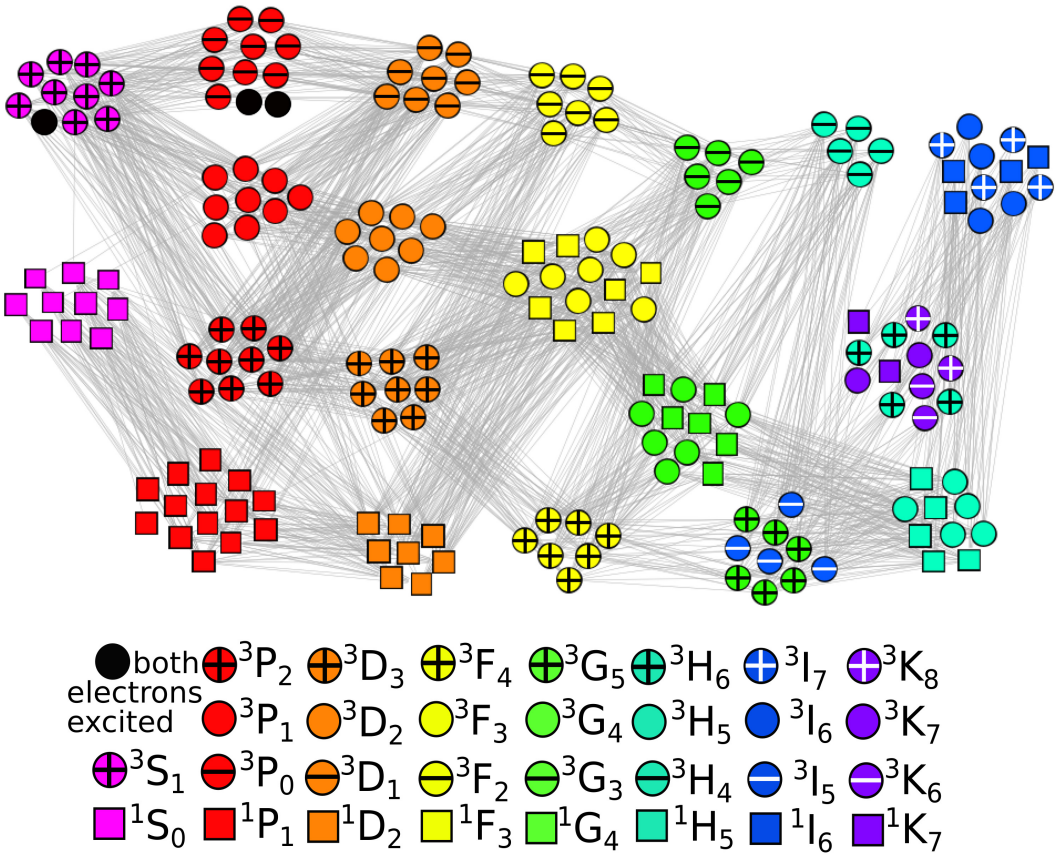


Figure 2.5: Helium network with positions indicating community membership; and color, symbol and shape indicating quantum numbers

triplet states with otherwise equal L and J (circles and squares in figure 2.5) are found to be in the same community.

In order to understand this the data used to create the helium network as described in 2.1 should be considered. For $L \geq 2$ all dipole transitions including singlet-triplet lines were calculated. Since weights are not included in this analysis, these transitions are treated to be as important as other transitions, even though they are significantly weaker. Therefore the singlet triplet splitting is invisible to the network and cannot be found.

The black nodes (doubly excited states) should be discarded from the further analysis, because they lie above the ionization threshold and are thus typically not found in nature.

Starting with $J \geq 5$, L and J are mixed in the communities. This happens due to the resolution limit of the NSBM: For helium groups with less than around $\ln 191 \approx 5$ nodes cannot be resolved. In this case it is interesting to analyze which quantum numbers the NSBM still finds. All communities even above this limit contain only nodes of the same parity. As discussed before the two parities correspond to bipartite sets of the network and are thus the most prominent structural feature. Furthermore, there are communities which break the L structure, but have same J ($^3G_5$ and $^3I_5$), as well as communities with different J but equal L ($^1I_6, ^3I_6$ and $^3I_7$). This indicates that the structure of L and J is complementary and with no difference in importance.

Iron

The communities found for iron are shown in figure 2.6. As stated above, the highest level of the hierarchical splitting corresponds to the parity. This shows that parity can be found even for complex system. Furthermore parity is put at the highest hierarchical order. As there is fundamentally no hierarchy in the quantum numbers, this can be interpreted as follows: In the network structure the parity corresponds to the straightforward feature of an almost bipartite network. This bipartite structure is easy to identify and thus assigned to the highest hierarchy level. This is further confirmed by the fact that no node has wrongly assigned parity. Thus even without weights the parity can be easily assigned.

The communities are also correlated with J, the second quantum number known for the description of complex atoms. This also covers both known selection rules: parity change and $\Delta J \leq 1$. In contrast to the parity there are deviations from the correlation between J and the communities. There are some communities, which

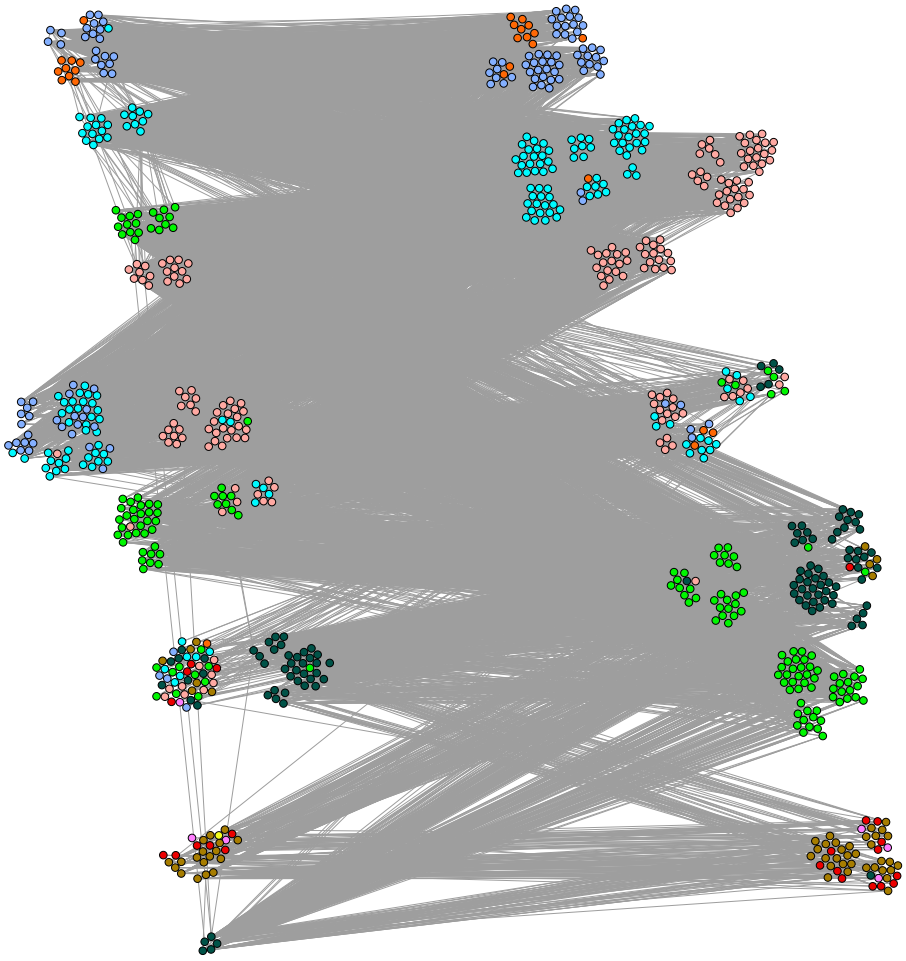


Figure 2.6: Iron network with colors indicating J and position indicating community hierarchy. The color codes are from $j = 0$ to $j = 9$: orange, blue, cyan, light red, bright green, dark green, brown, red, purple, yellow. The four level hierarchy is indicated by proximity: left-right is most cross-grained, small clusters are the finest level, intermediate levels are indicated by distance of clusters. Left-right is also even-odd.

contain single states that were assigned to a community dominated by a different J. For many of these states only few transitions are known, so that the NSBM has to work with limited data, which might lead to several errors.

Furthermore the network shows structure on top of parity and J. There are for example five communities with bright green nodes of the left side (even, $J = 4$). These communities belong to different assignments on higher hierarchy levels as well. The further splitting can be interpreted as follows: In addition to the splitting by J, there are further features in the data indicating whether a transition between two states is measured or not.

This additional structure is sometimes more prominent than the structure created by the $\Delta J \leq 1$ selection rule. This can be seen by the splitting of the green nodes into different groups even in the second hierarchy level. Simpler structure is related to higher hierarchical levels, so there should be a significant difference in the connectivity of the different nodes. This difference in connectivity corresponds to a difference in measured transitions in the physics picture. The interpretation of such additional structure is discussed in the next chapter at the example of thorium II. Here, this structure is related to the states energies and electron configurations.

Whether or not such additional structure is found is dependent on the amount of data available. In contrast to the green nodes, the orange nodes accumulate all in one community for each parity, and the red, purple and yellow nodes do not get their own communities at all. Considering the finite resolution limit discussed earlier, this result can be understood as a direct consequence of the different number of states known for different J.

This difference in the number of known states is also a likely explanation for the fact that no hierarchy level is found for J. Due to the resolution limit, if more states are known for a particular J, more structure can be found, which results in additional hierarchy levels. If instead less data is known for one J than for the others, no additional structure can be found for this J. On the higher hierarchy levels nodes with such a J might be sorted to a neighboring J. This happens for example for the orange $J = 0$ nodes, which are sorted to the blue $J = 1$ nodes.

One obvious error in the findings is that there is one community which contains a lot of different Js (left, third from bottom). All nodes in this community have very few links. The formation of such a community should be prevented by the degree correction build into the NSBM, which obviously fails here. No simple explanation for this was found.

The above described error sources all have in common, that they originate from the data and follow the philosophy, if there was more data, the results would be better. Though this is probably true, it has to be considered that the grouping found is even according to the NSBM probably not the best grouping, since in 100 tries this community structure was only found once. Further it might be that the base assumption of the NSBM just fails, since the connections between the communities are not statistical, but caused by quantum mechanics and choices of measurement, which have a high correlation amongst all the data.

Thorium II

In figure 2.7 the thorium spectrum is shown. It looks similar to iron in that the highest order splitting corresponds perfectly to parity and the lower orders correlate with J. Comparing to iron two main differences are found. There are fewer errors, especially on the left side (even parity), and the J splitting is clearly given by the intermediate level. The correspondence J intermediate level fails for high and low J, where less states are known.

Just like for iron, the lowest hierarchy level communities induce a finer splitting than parity and J. For example the left, orange group (even, $J = 3/2$) is split into three separate communities. Just like for iron this additional structure indicates that there is additional structure in the spectral data, which will be further analyzed in section 2.4.

A further difference between thorium and iron, which cannot be seen in the figure, is given by the relative likelihoods of different partitions. Of the 100 times the NSBM algorithm was run, for thorium 63 similarly likely (relative likelihood $< \exp(5) \approx 150$) different partitions have been found. These usually have a different assignment of the nodes for which J is not known. In contrast, the second most likely community structure found for iron is less than $\exp(5)$ times as likely as the best fit found. This indicates that, although the global minimum of the description length is hard to find for both atoms, the reasons for this might be different in both cases. For thorium some nodes are fundamentally hard to assign to a J based on the network structure, while for iron this structure is set and the finding of the global minimum is a computational problem based on too many nodes.

It can be seen that with standard community detection a meaningful splitting of nodes into groups is found. This splitting corresponds to the quantum numbers which are relevant for the selection rules: parity, L and J in helium, and parity and

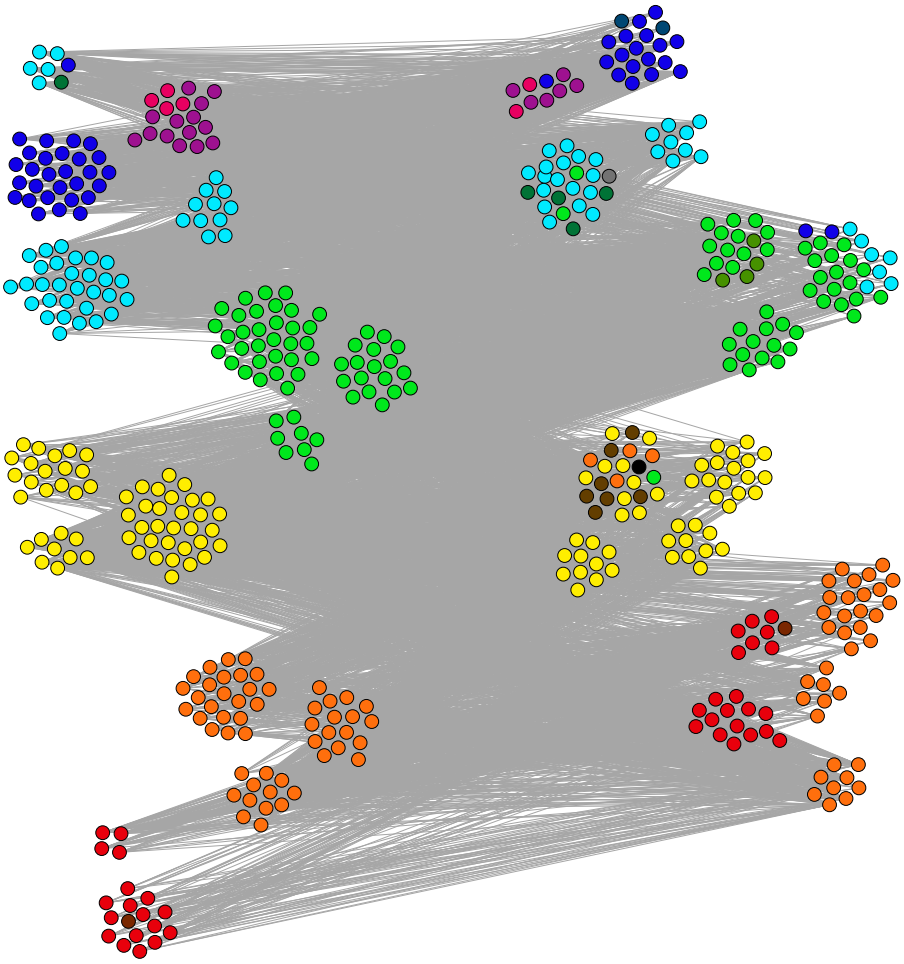


Figure 2.7: Thorium II network with colors indicating J and position indicating community hierarchy. Color code: J from 1/2 to 15/2 : red, orange, yellow, green, cyan, blue, purple, magenta; mixed colors correspond to unknown J. Three level hierarchy given by: left-right; big groups; close clusters. Left-right is also even-odd

J in iron and thorium II. The structure found this way is limited by the resolution limit of the NSBM or systematic features of the data like the lack of distinction between different S for $L \geq 2$ in helium. Furthermore, a hierarchy in the quantum numbers is found: the parity seems more important for the structure than J (and L). For iron and thorium II a finer splitting than the known quantum numbers is found.

2.3.3 Quantitative Evaluation

In order to evaluate the results quantitatively the adjusted rand index (ARI) will be used. The ARI is a number smaller than 1 to quantify the correlation of two partitions. It was introduced by Hubert and Arabie [1985]. It is normalized in a way such that an ARI of one corresponds to identical partitions and an ARI of zero corresponds to random correlation. Negative values mean less correlations than expected for two random partitions. The ARI is further symmetric under the exchange of the two partitions. It is calculated by checking whether the pairs of nodes, which are in the same group in one partition, are in the same group in the other partition as well.

Now the communities found by the NSBM are compared quantitatively to the different quantum numbers known for the elements. If one quantum number is not known for all states – as is the case for some thorium states – the respective state is for this quantum number excluded from the ARI calculation.

When talking about electron configuration we refer to a combination of orbitals (n and l) for the individual electrons. The use of these assumes that the total electronic wave-function is separable into single electron wave-functions. This is a good approximation for atoms with a small number of electrons, but not for atoms like iron and thorium. In this case, the electrons are entangled and a superposition of multiple configurations needs to be considered to describe one state. In the NIST database for each state the dominant components of the wave-function and their respective probability in the basis of single electron orbitals are given, including also the term. For example the ground state of thorium II is assigned as:

43%	$6d^2(^3F)7s$	$^4F_{3/2}$
27%	$6d7s^2$	$^2D_{3/2}$

Furthermore, a dominant component is assigned usually as the most likely one, but

for example for the thorium II ground state it is: $6d7s^2$. From this the configuration is chosen as just the combination of the electronic orbitals: For the 43 % component in the thorium II ground state this would be $6d^27s$. Thus the configuration $6d7s^2$ would be assigned to the thorium II ground state. For iron a term is assigned in the same manor in addition to the configuration. This term is given sometimes by jj-coupling and sometimes by LS-coupling. As the numbers for term and configuration stem from approximate theoretical calculations it is further not clear how accurate they are. Overall the choice of data for observables that do not commute with the Hamiltonian cannot and does not fully mirror the underlying physics.

Above we found that for iron and thorium the highest hierarchy level corresponds to parity and all lower levels are further separations of this. This correspondence is almost true for helium as well. Hence, none of the lower hierarchy levels can correspond to just one quantum number, but they can always only correspond to combinations of the parity with another quantum number. The first hierarchy level of thorium does not correspond to J, but to parity-J. Thus, only these combinations are analyzed. Sometimes the combination does not need to be explicit. For example each configuration can only belong to one parity: the product of the individual parities of the orbitals.

Helium

The ARI of the communities and the groups induced by the quantum numbers is shown in figure 2.8. Two of the four hierarchy levels have a good correspondence in physics: The highest hierarchy level corresponds to parity (ARI = 0.8), the lowest level corresponds to either LJ or SLJ (the term; both with ARI = 0.8). For the parity, only the group ${}^1S_0$ is assigned to the odd parity community. All other nodes are assigned correctly. The lowest level correspondences and the spin were discussed above. The apparently reasonable correspondences of spin and parity with level three (and two) are here just caused by the parity splitting. This can be seen as both ARIs shrink compared to the pure parity splitting.

As discussed earlier, there is usually no fundamental hierarchy between different quantum numbers. This explains that not every intermediate level has a clear physical meaning. Above the hierarchy was interpreted in a way that the easiest structure is found first, and then step by step all of the structure is uncovered. For helium it is not clear whether J or L is the more prominent structure, which can be seen in figure 2.8 by the fact that there is no correspondence for either of the intermediate

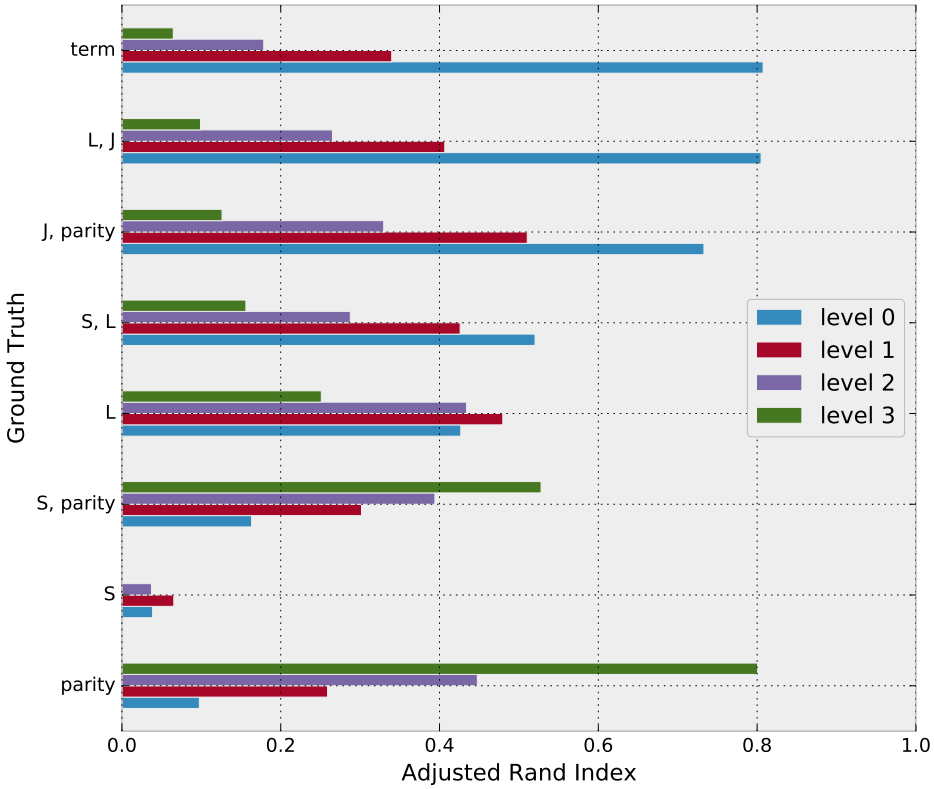


Figure 2.8: The ARI for helium comparing the NSBM to different ground truths.

levels.

Iron

Just like for helium, two hierarchy levels of iron have correspondences in the quantum numbers: The parity corresponds to level 3 with an ARI of 1, the combination parity-J corresponds to level 1 (ARI ≈ 0.65). As discussed above with figure 2.6 the correspondence of hierarchy level and J varies for different values of J, which explains the low value of 0.65.

All further choices of ground truth that were considered lead to only slightly better than random correspondence (ARI < 0.3 for almost all of them). Thus, similar to helium, level 2 is an intermediate level belonging to no obvious quantum number. For level 0, no correspondence is found among all quantum numbers considered. The best fit is found by comparing to J and the configuration, but this combination fits better to level 1 than to level 0. Thus, the additional structure found in the spectral data does not correspond to one of the considered observables, which are quantum numbers only for few electron systems.

Thorium II

For thorium for most states only parity, J and configuration are given in the NIST database. The term (LSJ for thorium) is only given for a few states. This makes analysis with respect to L and S impossible. Based on the results for iron they can be expected to be a bad fit for the communities though, as thorium II has more electrons than iron.

The main results are:

1. Parity corresponds perfectly to the highest hierarchy level.
2. J-parity corresponds well to the hierarchy level 1 (ARI $\simeq 0.7$).
3. Hierarchy level 0 has no good quantum number assigned.

As shown above in figure 2.7, for high and low J, J is only identified on the lowest hierarchy level. Thus the correspondence of parity-J to level 1 is only slightly better than that to level 0 (ARI ≈ 0.7 compared to ARI ≈ 0.5). For thorium it can also be seen that level 0 is the best fit for J-configuration. This indicates that the additional splitting found by the NSBM could be related to the configurations. This assumption is further analyzed in section 2.4.

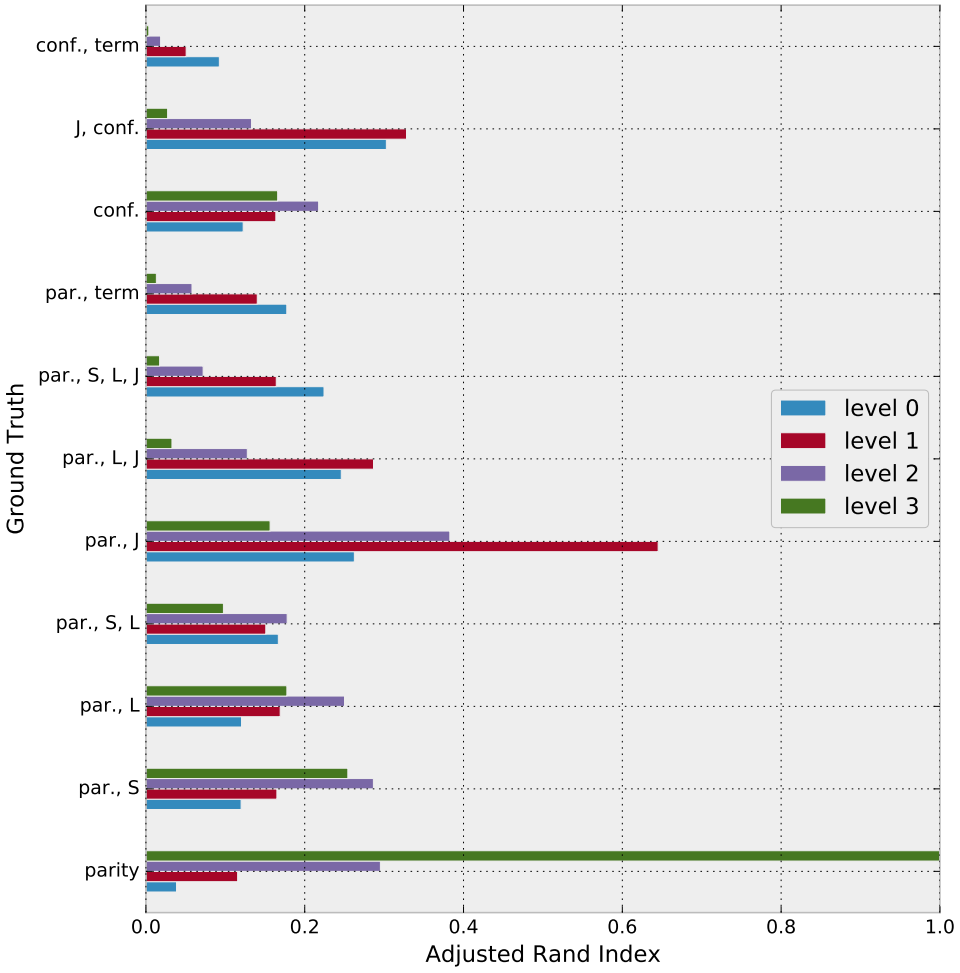


Figure 2.9: The ARI for iron comparing the NSBM to different ground truths. conf.: configuration; par.: parity.

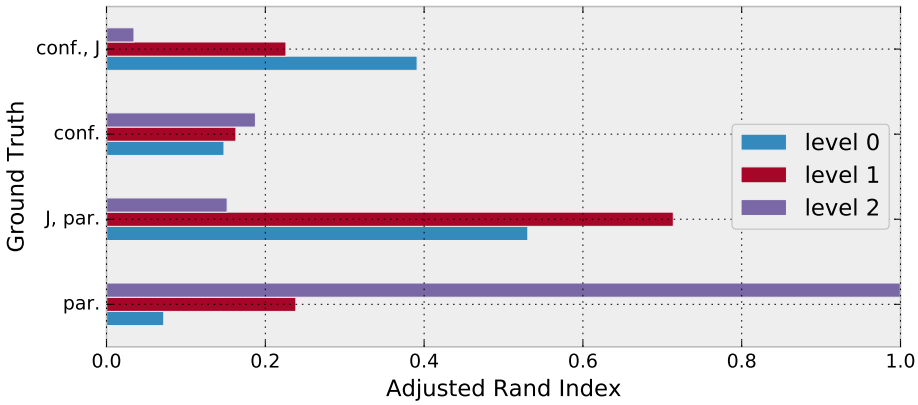


Figure 2.10: The ARI for thorium comparing the NSBM to different ground truths. conf.: configuration; par.: parity.

2.3.4 Conclusion

Community detection is a powerful tool in order to find structure in the spectrum and to identify similar states. Employing the NSBM without any further assumptions or modifications out-of-the-box, communities that correlate strongly with the known quantum numbers are detected. These quantum numbers are L, J and S for helium and the parity and J for iron and thorium.

In addition to the quantum mechanical structure, artifacts of the data choice are found. Only few states are known for extreme values of J or L, and thus the communities do not capture all the structure of the states for these values. For helium the discarding of the transition strength also leads to a loss of differentiation between singlet and triplet.

To further improve the results and find the structure here, different approaches can be taken. In order to tackle the singlet-triplet mixing of helium the weights of the transitions could be taken into account. A straightforward way would be to introduce a weight cutoff that is high enough to cut most of the transitions between singlet and triplet. Afterwards this structure should be found again. In general such a weight threshold is arbitrary though, so more sophisticated methods like a weighted version of the SBM as described by Clauset et al. [2013] and Peixoto [2018a] could be used.

Furthermore, different methods that are specifically tuned to describe atoms could be developed. One approach is to specifically look for disassortative structure, as proposed by Lackner et al. [2018]. This method turns out to capture the structure

of helium better than the NSBM, but it performs worse for iron and thorium. Here, further approaches could be made considering further structure, which is known about atoms. One example would be to take into account that quantum numbers have no hierarchy, but there are different possible splittings and these capture different structural features. This would correspond to finding different community splittings on the network side.

For both iron and thorium, in addition to the known quantum numbers parity and J, a finer splitting was found. This indicates that there is structure in the data that is not captured by these quantum numbers. As the NSBM is tuned to only find statistically relevant communities, it is unlikely that these additional communities are caused by statistical fluctuations in the data. Therefore, these communities can have two causes: measurement artifacts or fundamental physics. On the one hand, measurements of the spectrum are not treating all lines equally, but usually focus on certain wavelengths. On the other hand only little is known about the physics of complex atoms. Hence, there could be thus far unknown structure in the atom. In the next section these communities will be addressed.

All in all, community detection offers a simple way to uncover the structure of the spectrum. This has been confirmed for simple atoms like helium, where essentially all known structure was recovered. For complex atoms, in addition to the known structure further communities are found in the spectrum.

2.4 Further Structure using Energies

In the last section communities were found that have no obvious correspondence to one quantum number. In order to analyze these communities, the energies of the states will be taken into account. The analysis with respect to energies will be done at the example of the thorium II ion, which will show a clustering of states in the energy domain. Furthermore the configurations will be analyzed in more detail, as some structure was found in the previous section relating the communities to configurations. In the end we will consider the wavelengths of the transitions and discuss whether we find a measurement artifact or physical structure.

2.4.1 Structure in the Energy

In figure 2.11 for each state its community is plotted against its energy. It can be seen that the level 0 communities form clusters in the energy domain. The level 1

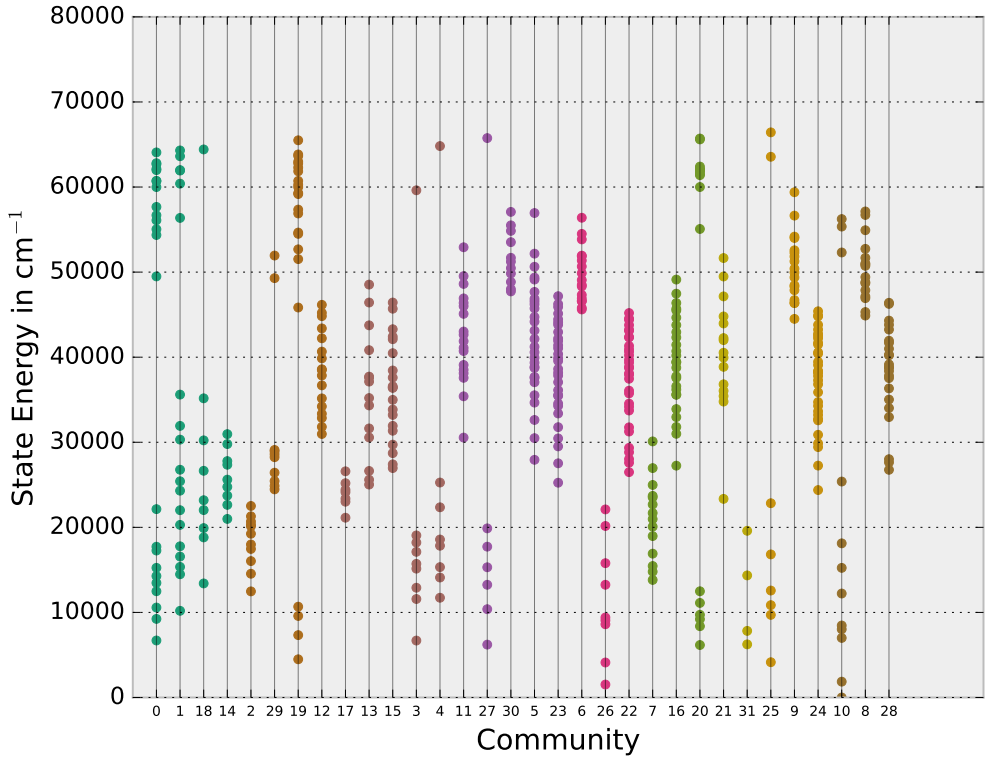


Figure 2.11: Energies of the states (points) in the communities. Equal color means equal level 1 community.

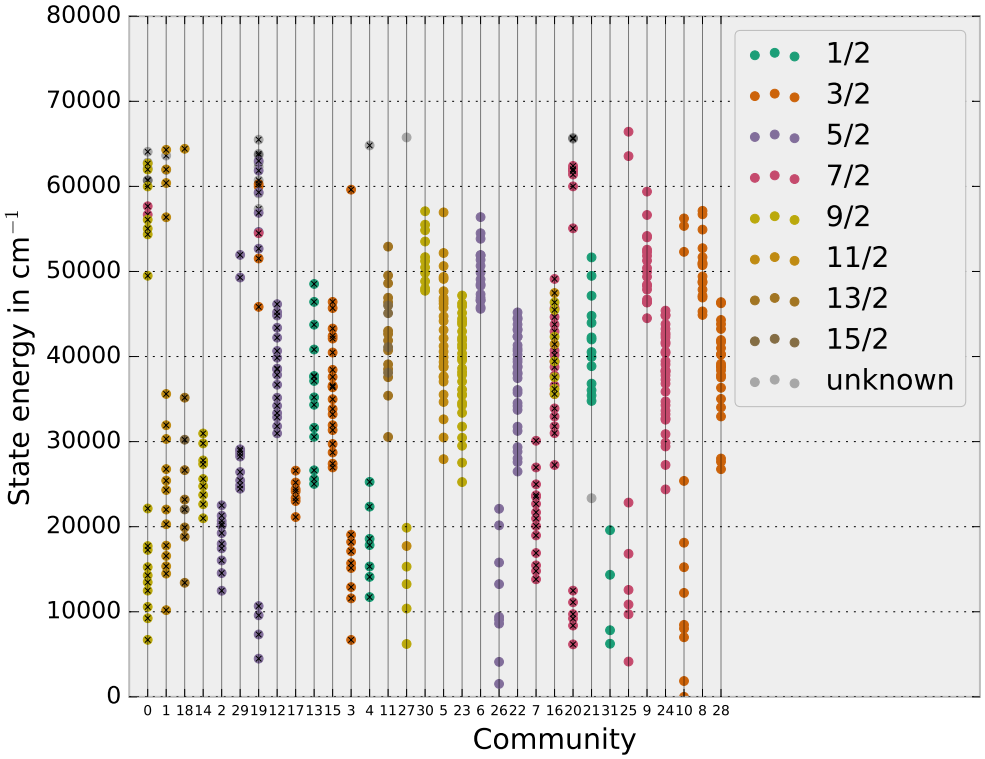


Figure 2.12: Energies of the states plotted by community. The colors encode J. An x indicates odd parity.

communities on the other hand do not have this property. Most level 0 communities correspond to one energy cluster, but some level 0 communities contain two different clusters. An example for this is given by community 0, which contains two distinctive energy clusters, ranging from $5000\,\mathrm{cm}^{-1}$ to $25\,000\,\mathrm{cm}^{-1}$ and from $50\,000\,\mathrm{cm}^{-1}$ to $65\,000\,\mathrm{cm}^{-1}$.

The clustering roughly corresponds to cutting the level 1 groups at energies of about $25\,000\,\mathrm{cm}^{-1}$ and $45\,000\,\mathrm{cm}^{-1}$. Only in a few cases multiple clusters in the same range form within one level 1 community. In figure 2.12 this is traced back to different parity-J groups in the same level 1 community, as can be seen for example for the communities 0, 1 and 18 completely to the left. Two different level 0 communities with same parity and J have no overlap in the energy domain except for individual states.

The results can be summed up as follows: The measured transition structure of the thorium II spectrum cannot be fully explained by the quantum numbers parity

and J. The additional structure is caused by cutting the states at certain energies: Different communities with the equal parity and J cover different energy ranges with almost no overlap. These cuts happen at the same energies for all different parity-J combinations: $25\,000\,\mathrm{cm}^{-1}$ and $45\,000\,\mathrm{cm}^{-1}$ for all even and most odd states. Thus, there should be some effect that results in groups of nodes having similar interactions with light also having similar energies. Some odd parity communities also contain two different clusters. This leads to the following questions:

1. Why do the communities correlate with energies?

2. What is the meaning of different energy clusters within one community?

The section is closed by a discussion if this structure can be attributed to physics or to measurements, and a further test to decide this is proposed.

2.4.2 Configurations in Different Communities

In order to analyze this structure the configurations are taken into account. In the last section in figure 2.10 it was already shown that the configurations are described better by the zeroth hierarchy level than by the first level. The correlation was low, but as only dominant configurations were considered, which are only a rough description of the electronic structure a large correlation was not expected.

In figure 2.13 the two dominant configurations are shown for each state of thorium II. For each node the fraction of the pie covered in one color corresponds to the probability of finding the state in the respective configuration. The white part is the fraction not covered by the NIST data. It could correspond to any configuration compatible with the respective parity and J, including the ones already shown in the node. The white nodes are the states for which no configuration was calculated.

First, this graph confirms that the configuration is not a good description of the state, as most of the nodes have two different colors. This means that the configurations of the two main components of the wave functions are different.

Despite that, a correlation between the configurations and the communities can be seen. Take for example the top right group in figure 2.13. Most of the nodes in community 13 belong to the two green configurations, whereas the nodes in community 4 belong to the purple configurations. These results generalize to many other communities in the top, for example with community 15 versus community 3 and 17, community 12 versus community 2 and 29 and so on. This is similar for the bottom

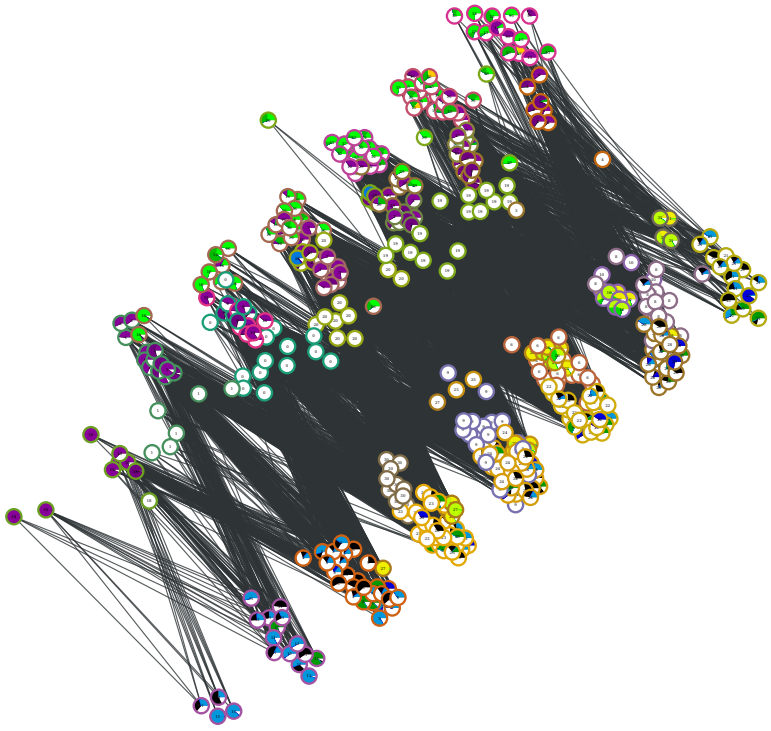


Figure 2.13: Configurations of the different nodes in the thorium network. The top groups have odd parity and the bottom groups have even parity. The further splitting is right to left by increasing J. The circle color around the node and the number within the node correspond to the community number, the pie charts within each circle give the contribution of each configuration. Similar colors do **not** mean similar configurations.

group. For example in the bottom right, community 21 corresponds to black and blue configurations, whereas group 31 corresponds to yellow-green configurations. The splitting into configurations is not perfect though, as for example also some of the states in group 13 have the same configuration as some states in group 4.

Furthermore, it can be noted that for the different Js the same configurations are found and split into the same communities. For example all top communities named above correspond to the splitting into purple and green.

One possible way to interpret the results is as follows: In addition to parity and J, there is further physical structure in the states, which influences their interaction with light. This physical structure is related to the configurations, so that states with equal configurations contributing to them have similar spectral properties. This further structure is then found by the NSBM. As the configurations have a strong correlation with the energies of the state, the energies of the states in the same community form clusters.

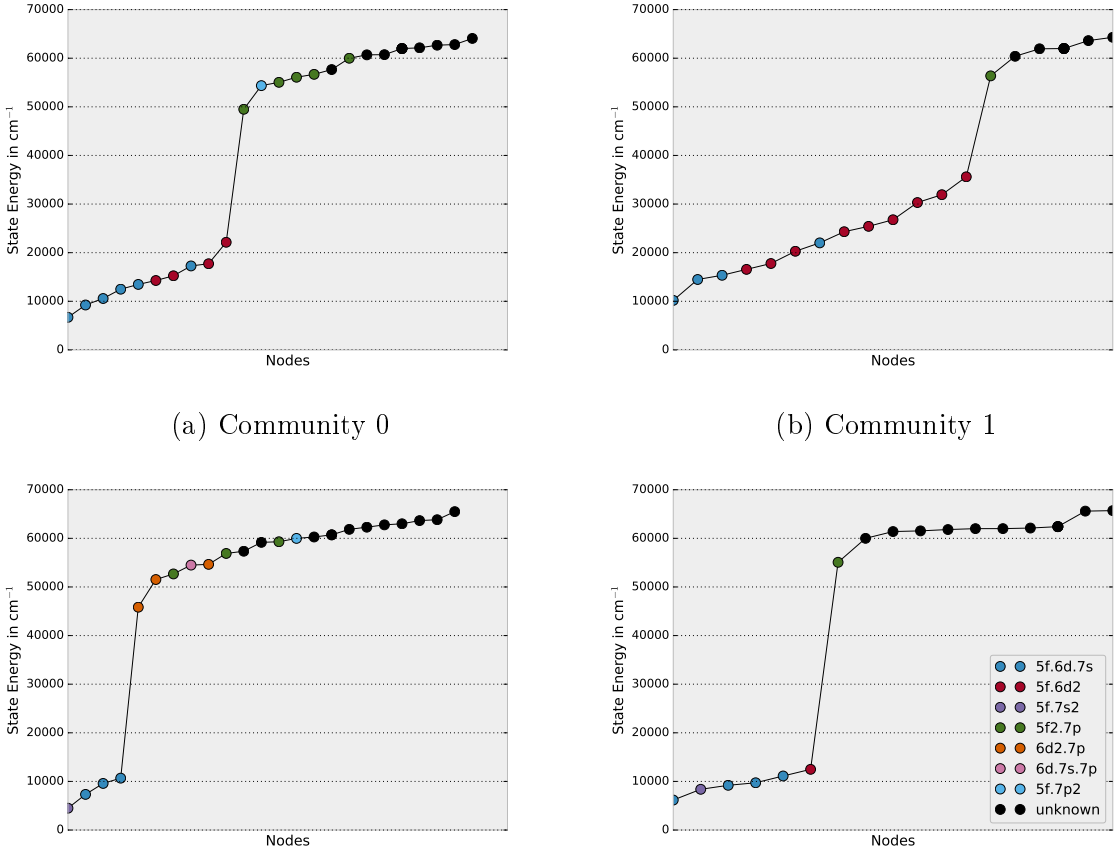
2.4.3 Configurations within one Community

Next, the communities with different energy clusters are analyzed with respect to their configurations. For simplicity, only the configuration of the dominant component of the wave-function as assigned by the NIST is used.

In order to analyze the meaning of the different energy clusters within one community, the energies of the individual states are compared to their configuration. This is done in figure 2.14. The x-axis is given by an ordering of the nodes by increasing energy, the color corresponds to the dominant configuration as given in the NIST database.

It can be seen that the states with low and with high energy tend to have different configurations from each other. There is a qualitative feature distinguishing the bottom from the top cluster: The top cluster contains states with the $7p$-orbital. As this is a high energy orbital, the contribution of this orbital to the state can lead to a jump in energy. Consider community 0 in figure 2.14a. Here, the low energy states are given by the configurations $5f6d7s$ and $5f6d^2$, whereas the known configurations of the high energy states are given by $5f^27p$ and $5f7p^2$.

This example further shows that the dominant configurations within each cluster vary as well. As shown in figure 2.13, multiple configurations need to be considered for one state. Among the different states the dominant configuration can switch from one configuration to another. Thus even for similar states, different dominant



(a) Community 0

(b) Community 1

(c) Community 19

(d) Community 20

Figure 2.14: States within the same community are plotted against their energies. Colors correspond to the configurations.

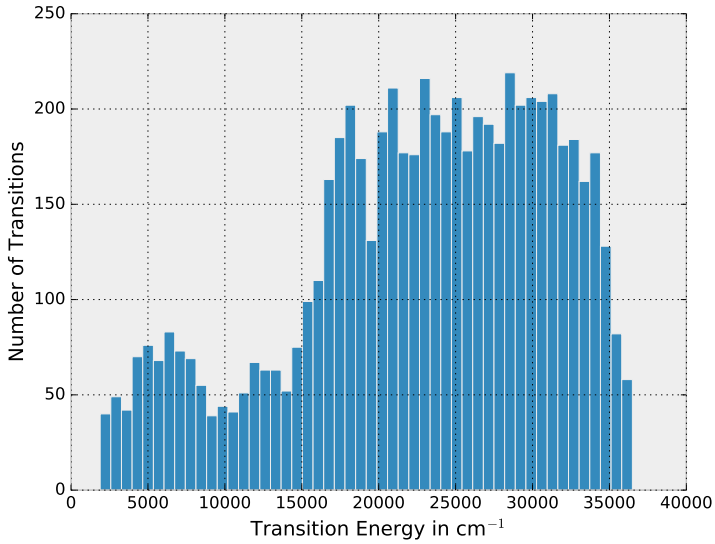


Figure 2.15: Thorium II transition energies.

configurations are expected.

2.4.4 Energy Spectrum of the Transitions

Now we turn to another energy related feature: the energies of the transitions involved. This analysis should give some intuition of the experimental bias caused by the measuring of certain wavelength ranges and its relation to further potential structure.

The transition energies of thorium are plotted in figure 2.15. The spectrum has the following structure: One block with few transitions ranges from 1700 cm^{-1} to 16 000 cm^{-1} and one block with many transitions ranges from 16 000 cm^{-1} to 36 000 cm^{-1}. To explain the structure the measurements made for thorium need to be considered. An overview is given by Redman et al. [2014]: Different measurements have been done, overall spanning the energies from 1700 cm^{-1} to 36 000 cm^{-1} The sharp transition at around 15 000 cm^{-1} can be explained by a measurement by Lovis and Pepe [2007], who measured transitions between 14 500 cm^{-1} and 26 000 cm^{-1}.

Looking at figure 2.12, the structure of the energy usually looks as follows: One cluster ranges from 25 000 cm^{-1} to 45 000 cm^{-1}, one cluster lies above 45 000 cm^{-1} and one cluster lies beneath 25 000 cm^{-1}. This relates to the transition energies in the following way: Nodes in the high and low energy groups are usually connected to nodes in the mid energy group. These transitions usually fall in the range of

$16\,000\,\mathrm{cm}^{-1}$ to $36\,000\,\mathrm{cm}^{-1}$. The connections top-top, mid-mid and bottom-bottom fall into the range $1700\,\mathrm{cm}^{-1}$ to $16\,000\,\mathrm{cm}^{-1}$. Top-bottom connections are (almost) never measured. This indicates that the structure beyond parity and J is not necessarily related to underlying physical laws, but could also be a pure measurement artifact.

In order to get further insight the transition energy histogram and community versus energy diagram have been plotted for other elements as well and can be found in appendix A. Here a similar structure with respect to the energies can be seen for many elements like thorium I (figure A.2), but none of the elements have a similarly structure for the transition energies. This indicates that the energy clusters are more than a measurement artifact.

Further tests need to be made before final conclusions can be made. One such test could be deleting transitions with high energies and see how this changes the communities. If the communities get smaller, this is an indication that the communities are caused by the measurements, otherwise they are likely a feature of the underlying physics.

2.4.5 Conclusion

We have shown that the communities found by the NSBM not only find the structure of the states corresponding to parity and J, but also further split them according to their energy. This splitting by energy corresponds roughly to a splitting into different groups according to their configuration, so that different communities contain different sets of configurations. Furthermore, it also relates to the spectrum of transition energies, which are mostly measured between $16\,000\,\mathrm{cm}^{-1}$ and $36\,000\,\mathrm{cm}^{-1}$ (278 nm to 625 nm).

The question whether this is a measurement artifact or actual structure cannot be answered conclusively, as there are indicators for both directions: On the one hand energy groups tend to have a size such that the wavelength peak, which was probably caused by measurements, fits roughly such that only transitions from one group to its neighbor in energy are possible. On the other hand we find some communities which cannot be explained by this and furthermore, a comparison between the groups and the communities shows that the underlying structure is correlated with the configurations. Furthermore, in appendix A it can be seen that the community structure of thorium I is similar to the one of thorium II, but the wavelength histogram looks different. To make a conclusive decision some further

tests should be made.

Moreover, we found that some communities have two clusters in the energy domain. These clusters differ by the fact that the configurations of the low energy nodes do not contain $7p$ orbitals, whereas the ones with high energies do. As this orbital has a high energy, it causes a splitting into two separate clusters. This can also be used to further separate these communities by whether the configuration has a $7p$ part or whether it has not.

We have thus shown that the structure found in the network beyond parity and J corresponds to cutting the groups with equal parity and J at the energies of about $25\,000\,\mathrm{cm}^{-1}$ and $45\,000\,\mathrm{cm}^{-1}$. In addition, this splitting also corresponds to differences in the electron configurations. Furthermore, some communities also have multiple clusters in the energy domain, which corresponds to a qualitative difference in the configuration caused by the high energy $7p$ state. These results will be used in the next chapter to improve the prediction of the configuration for individual states using network science.

2.5 Predicting Quantum Numbers from Network Structure

In the previous two sections 2.3 and 2.4 groups of nodes that have similar quantum mechanical properties were discussed. In this section we now want to take this one step further and go from identifying properties to predicting them. To see the relevance of this problem, we can take another look at figure 2.7. Here many nodes are colored dark because their J is unknown, and there are even fewer known configurations. In order to get a better theoretical understanding of thorium II (and maybe even to predict new energy states), it would be very helpful to know the values of the quantum numbers and other properties for the individual states.

Thus, in this chapter we will propose three different methods to predict these values and compare them for hydrogen, helium, iron and thorium II. In general the following setting is considered: The properties of one specific state will be predicted, given we know these properties for all other states.

2.5.1 Methods

Groups

First, the communities detected by the NSBM will be used for the prediction. In section 2.3 these communities were found to correlate well with various quantum numbers. This motivates the following scheme:

1. Find communities according to the link structure (via NSBM)
2. Choose the community of the unknown node
3. For all other nodes in this community add one vote to the respective quantum numbers of the nodes
4. Repeat steps 1 to 3 and weight the votes with the likelihood assigned to the communities
5. The quantum numbers with the highest votes are the guess for this node

The sampling here is the preferred method from a Bayesian statistics point of view Peixoto [2017], as we might encounter a situation, in which the most likely community structure proposes one quantum number, but many others propose a different one, which is thus overall more likely.

Energies and Groups

In section 2.4 we found that some of the communities identified by the NSBM contain different clusters in the energy dimension. One way to interpret this result is that these communities should be two separate ones. This assumption is supported by the result that the nodes in the different subgroups typically belong to different sets of configurations. This leads now to a slightly different algorithm:

1. Find the best community structure (as ranked by likelihood)

2. Split the communities, which contain different energies, into different sub-communities

3. Do a majority voting within the (sub-)community of the unknown node to assign quantum numbers

With this method it cannot be sampled reasonably, as there are no likelihoods known for communities split according to energies. Thus we will only look at the best proposed group.

Eigenvectors

Next a completely different method will be proposed. It is motivated by the finding that the eigenvectors of the adjacency matrix (defined in equation 2.1) are a good indicator for the structure of the network (see for example Chung and Graham [1997]).

From the definition it is clear that each row and column of the adjacency matrix correspond to one node of the network, and accordingly each element of an eigenvector corresponds to one node of the network. This way for each node a feature vector can be created by taking the components of the (normalized) eigenvectors belonging to this node. In order to encode the importance of each eigenvector, they are scaled by the square of the respective eigenvalue. Thus, the feature vector $\vec{u}^{(i)}$ of node i is given by:

$$u_j^{(i)} = \lambda_j^2 v_i^{(j)} \tag{2.5}$$

λ_j is the jth eigenvalue and $v_i^{(j)}$ is the ith component of the jth eigenvector.

Next, the distance d_{ij} of two nodes i and j can be assigned as the distance of the respective feature vectors:

$$d_{ij} = |\vec{u}^{(i)} - \vec{u}^{(j)}| \tag{2.6}$$

The similarity s_{ij} of the two nodes is now defined as:

$$s_{ij} = \exp\left(\frac{d_{ij}^2}{8 \cdot nd(i)^2}\right) \tag{2.7}$$

$nd(i)$ is the distance from node i to his nearest neighbor in feature space. The specific scalings (squaring the eigenvalues, using a Gaussian cutoff with $2 * nd(i)$ as standard deviation) are motivated by showing good results in tests. The prediction is now given by the quantum number with the maximum score, which is defined by:

$$\text{score}(n^*, i) = \sum_{\text{nodes } j} s_{ij}\delta(n^*, n_j) \tag{2.8}$$

n^* is the quantum number for which we calculate the score, n_j is the respective quantum number for node j and $\delta(n^*, n_j)$ is 1 for $n^* = n_j$ and 0 otherwise.

2.5.2 Results

Next, the quality of the prediction will be analyzed by the prediction accuracy, which we calculate as follows: We go through the nodes of the network one by one, delete the data of this particular node, predict it again using the methods described above and then compare the predicted quantum numbers to the actual quantum numbers. The fraction of correct predictions for each quantum number and each atom can be found in the figures 2.16, 2.17, 2.18 and 2.19.

Let us first look at the community detection methods (labeled "groups" in the figures). For parity, which has the simplest selection rule, the methods works perfectly. The same is true for L in hydrogen and helium. For hydrogen, the prediction of J and term is comparatively hard. The reason is probably that groups with a single term (3 nodes for large L) are too small for the system to be detected, and only the angular momentum can be identified (L and J would be enough to determine the term for hydrogen). For the other elements analyzed (helium, iron and thorium) J can be predicted very well.

For helium, the prediction of the combination of spin and angular momentum and

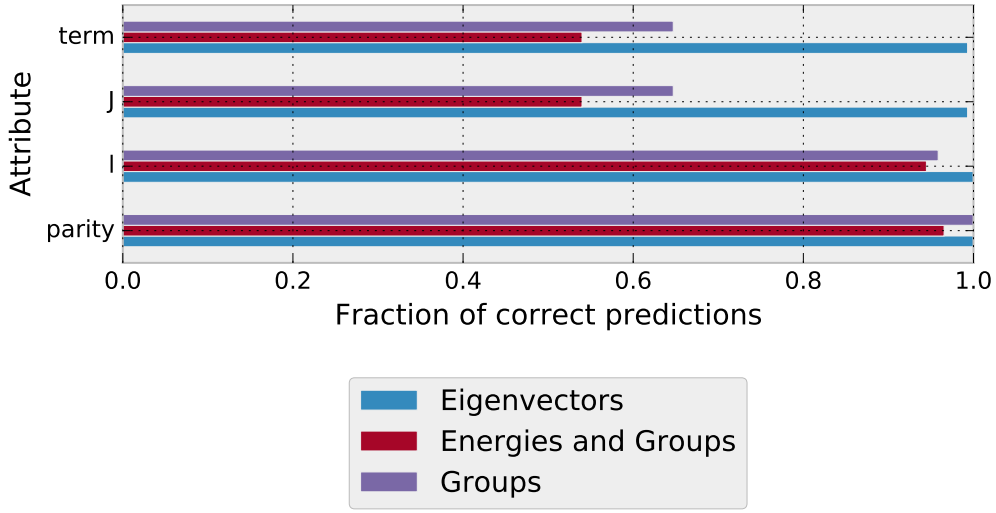


Figure 2.16: Fraction of correctly predicted attributes in hydrogen by prediction method.

the prediction of the term are more difficult. This is caused by the difficulty to find the spin in the network (compare section 2.3).

For iron we can see that the term cannot be predicted at all (about 10%), but the electron configuration is somewhat predictable (about 50%). For thorium the term is not given in most cases and therefore not analyzed. For the configuration we find similarly to iron about 50% accuracy. It should be noted that neither configuration nor term are good quantum numbers describing the wave function. Instead they are approximations given by analytical calculations and describe the state up to some degree of accuracy.

Taking the energies into account changes the prediction accuracy only very little. The only significant differences are: For hydrogen the prediction accuracy of term and J drops from about 65% to about 55%, and for iron the prediction accuracy of the configuration increases to roughly 60% and for term it increases to 30%. As for hydrogen the energies are just given by the Rydberg formula, no clustering should appear and thus the energy method should not work. For iron configuration and term correlate with energy, thus differentiating by energy increases the prediction accuracy.

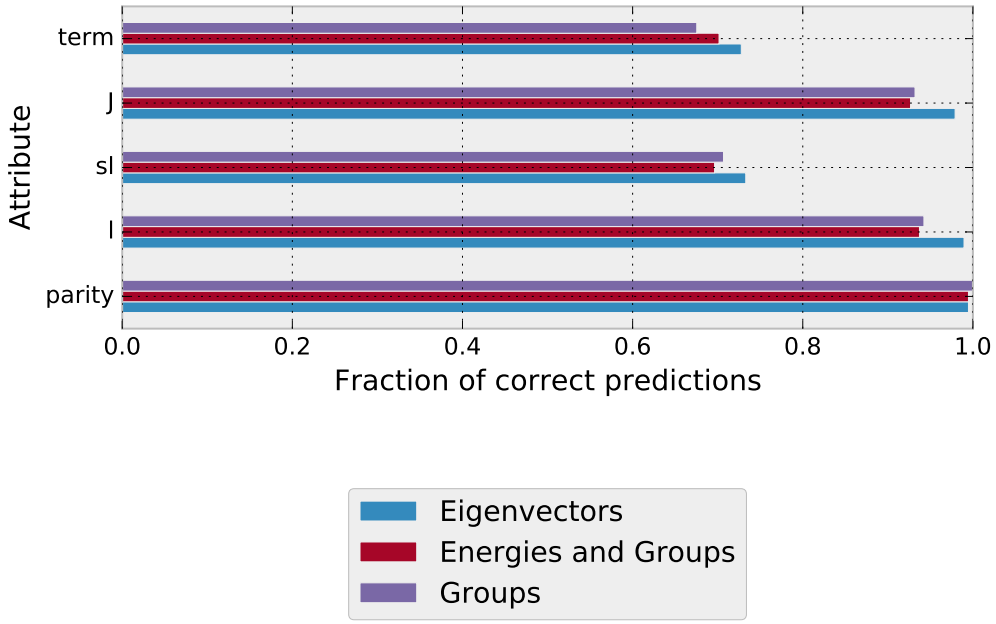


Figure 2.17: Fraction of correctly predicted attributes in helium by prediction method.

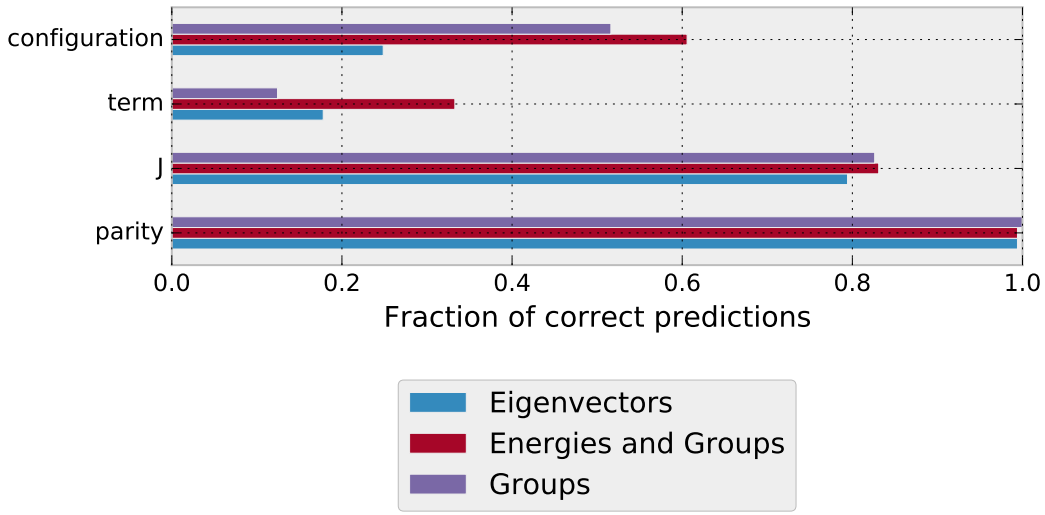


Figure 2.18: Fraction of correctly predicted attributes in iron by prediction method.

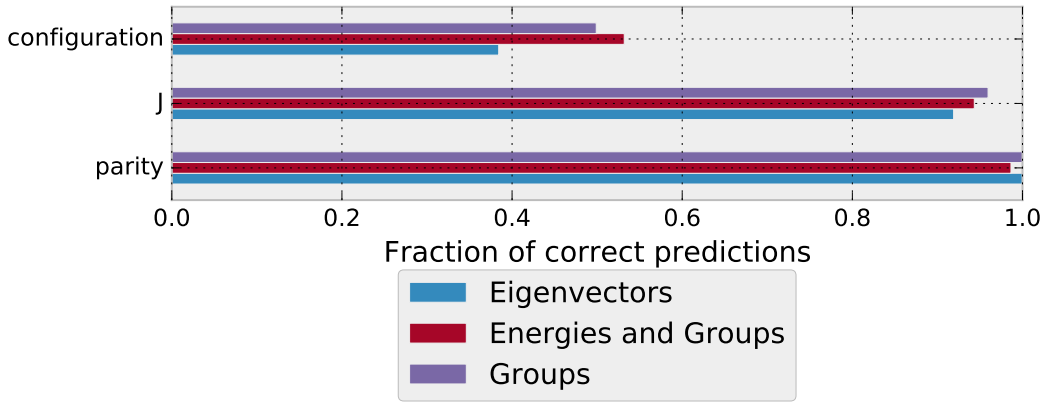


Figure 2.19: Fraction of correctly predicted attributes in thorium by prediction method.

Considering the eigenvectors instead of communities for prediction works almost perfectly for hydrogen. For helium, similar as with the communities, problems arise when predicting quantum numbers that contain the spin. For iron and thorium configuration and term are not predictable, for iron both being around 20% in accuracy, and for thorium the configuration can be predicted correctly in 40% of the cases. Parity and J can be predicted well in all cases.

2.5.3 Conclusion

Overall we find that all three methods proposed are very similar, with slightly different strength for the different special cases. The quantum numbers that describe the eigenstates very accurately (parity and J, as well as L and term for hydrogen and helium) can be predicted well. The special selection of data for the hydrogen spectrum (very few states with very high angular momentum) makes it hard to predict J – and thus term – via communities, most likely because these communities would fall under the resolution limit. The eigenvector method on the other hand seems to not care about this as much, as it does not rely on finding communities.

Finding properties that are not quantum numbers (term and configuration for iron and thorium) works significantly worse. Determining the configuration works also better than determining the term and in order to get the best results, energies should be taken into account. This happens because these quantum numbers have a large effect on the states' energies. On the other hand the eigenvectors fail here completely. To interpret this we have to consider that both of these assignments are not good quantum numbers, meaning that their assignment is not unique and that each state can have contributions of for example multiple configurations. Further these quantum numbers have less impact on the spectroscopic structure, but more impact on the energies, leading to the better accuracy when using energies.

All in all the quantum numbers, which correspond to known symmetries of the Hamiltonian and are relevant for the spectrum, can be predicted well, and all methods show good results in most cases. The properties that do not correspond to symmetries of the Hamiltonian are harder to predict, but can still be predicted to some extend if the state energies are taken into account in addition to the network structure. Thus, the network can be used to predict quantum mechanical properties of single states, especially if they are well defined.

3 Predicting Transitions and Levels

In the previous chapter, network communities were analyzed with respect to quantum mechanics. In this chapter however, the network is used directly to predict new transitions and energy levels. The results can then be compared to the physical ground truth and thus allow the benchmarking of various methods. This chapter is split into two parts: in section 3.1 transitions will be predicted, and in section 3.2 new energy levels will be considered. For the latter problem, only the transitions of the new levels will be predicted.

As in any physical theory, predictions which can be compared to measurements need to be made. This comparison allows us to disprove a model or find the limits in which a model is valid. Furthermore, different network methods can be benchmarked against one another, thus finding which properties work best in which cases. For this kind of prediction, transitions are clearly the go-to-property to predict because they are easy to measure. In fact, all that needs to be done is take a picture of the spectrum and then the peaks in that picture need to be identified. In our case, this corresponds to checking whether there is a peak at a certain position or not. In contrast, quantum numbers can only be accessed by complicated laser setups – even though they are by definition observables of the system. Hence, most of the known quantum numbers were not assigned by measurements, but by comparison to a theory which relates quantum numbers to energy levels, or they were directly inferred from the selection rules.

Predicting new transitions and states can also have practical relevance: in order to use the spectrum as a fingerprint of the atom, it is important to have as much information as possible about said spectrum. Although measuring individual transitions is easy, checking all transitions allowed by selection rules is a tedious task. Here, a prediction of specific transitions would decrease this effort significantly: a likelihood is assigned to the individual possible lines and they can then be checked in order of decreasing likelihood, thus significantly reducing the number of non-existent lines examined. Knowledge of these new transitions and states can further be used not only as a fingerprint of the atom, but also to extract new information about the

atomic structure.

This kind of prediction is impossible for state of the art methods in numerical quantum mechanics, as analyzing the behavior of interacting and thus entangled particles is very difficult. Thus new, empirical approaches like the one proposed in this chapter might shine.

3.1 Link Prediction

First we are going to tackle the task of finding transitions between two known states. This translates to a well analyzed question in network theory: "The link prediction problem". This problem can be formulated as: "Can the network structure be used in order to predict further existing links in the network?" It is topic of ongoing research. Solving this problem also leads to very practical applications: Recommending new products in online shopping portals or proposing new friends in social networks are both done by link prediction. Many algorithms have been developed to approach this problem, which take the network as an input and output a ranking of all possible links by their likelihood of existence.

The various techniques developed in this field have different strength and weaknesses. A review can be found by Lü and Zhou [2011]. Many methods use the local structure of the network, as this scales well to large networks. There are however also many algorithms taking the global structure in account. In this section we focus on the latter type, as it has proven to work well on spectroscopic networks and the computational complexity is not a major concern for these networks with up to a few hundred nodes.

As no new measurements are available, a random dropout will be employed in order to evaluate the prediction accuracy. The general scheme is explained in figure 3.1: First some links (red) are removed from the measured network (figure 3.1a). Then the modified network is used for prediction (figure 3.1b). This leads to a ranking of links, which can then be compared to the original network (figures 3.1c and 3.1d). Links which were deleted from the original network in figure 3.1a are considered as true links or correct predictions, links which were not part of the original network are considered to be false links.

This chapter is a short summary on the work conducted in this area in collaboration with Julian Heiss and Armin Kekić. For a more elaborate discussion please refer to the master's thesis of Julian Heiss.

3.1.1 Methods and Observation

First, the methods used for link prediction are introduced. Two methods which have been found to work well will be analyzed: The nested stochastic block model and the structural perturbation method. For a comparison various other methods please refer to the thesis of Julian Heiss.

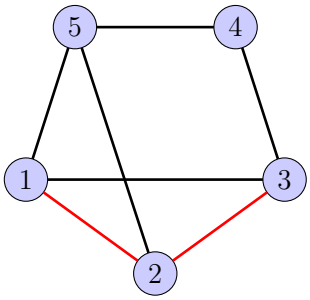
In order to evaluate the data the Receiver Operating Characteristic curve (ROC curve) will be used (Hanley and McNeil [1982]). It is shown schematically in figure 3.1d: The links predicted are analyzed in order of decreasing likelihood. For each predicted link, the fraction of recovered correct links with likelihood larger than this link (*true positive rate*) is given by the *y-coordinate*, whereas the fraction of recovered incorrect links with larger likelihood (*false positive rate*) is given by the *x-coordinate*. Another interpretation is given by: Each true prediction is one step up and each false prediction one step to the right, starting from (0, 0) and going to (1, 1). The ROC curve of a perfect prediction is the Heaviside step function, whereas a random prediction induces a diagonal ROC curve: $y = x$

The further formulation are used:

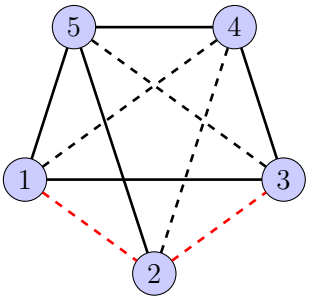
- The *ground truth* is the hidden, underlying structure of a system. In the dropout scenario ground truth also refers to the network before dropout.
- A *true* link is a (potential) link that exists link in the ground truth.
- A *false* link is a (potential) link that does not exist in the ground truth.

Nested Stochastic Block Model

The NSBM was already introduced in section 2.3 for community detection. It can also be used for link prediction. The method is described by Peixoto [2017] and references therein; specific information for how it was used is given in the appendix B. The NSBM partitions the nodes into communities and gives for each pair of communities A and B a number of links e_{AB} connecting these two communities. This number can be converted into a probability p_{AB} given as the fraction of observed edges over potential edges. p_{AB} can be interpreted as the probability of any two nodes n_1 and n_2 in communities A and B respectively to be connected. The community knowledge can now be used to determine the likelihood of a link n_1 and n_2. The naïve assumption p_{AB} for this likelihood is further modified by the degrees of

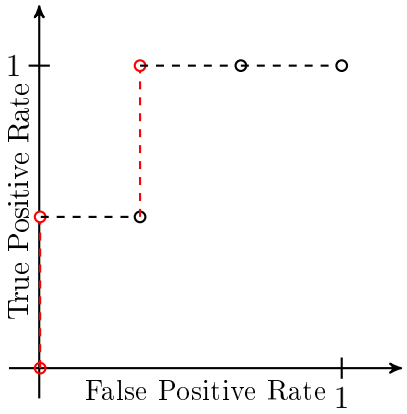


(a) True network (before dropout)



(b) Prediction is done

(c) Edges sorted by likelihood



(d) ROC is plotted

Figure 3.1: Blueprint of the dropout evaluation used in this chapter: 3.1a shows the network according to the ground truth. The red links indicate data which is dropped out. 3.1b and 3.1c show the prediction algorithm. For all potential edges a likelihood is assigned and the edges are sorted according to that likelihood. In 3.1d the creation of the ROC is shown: Going from top to bottom through the links in 3.1c, each true prediction translates to one step up and each false prediction to one step right.

the two nodes and by the already known structure of the network. This cannot be done by exact calculation, but approximative methods need to be employed. For each link a probability of existence is acquired, which can later be compared to measurements.

The ROC curves for different atoms are shown in figures 3.2 to 3.4. For all three atoms almost all true links are ranked higher than 80 % of the false links. For helium the line rises almost diagonally from (0, 0) to (0.2, 1). Thus, the remaining 20 % false and all true links are ranked with no clear preference towards the true links. For iron and thorium the ROC curve starts off very steep and then flattens out. This indicates that contrary to helium here the true links rank on average a lot higher than the false links. This also means that the ROC curve of thorium and iron – and to a lesser extent the one of helium as well – is very steep in the beginning. Therefore almost all of the high ranking links are actually part of the ground truth. Because of this feature it is possible to use the NSBM prediction to guide further measurements, as among the high ranking links many transitions should be found. This can also be used for an iterative procedure, modifying the network used for link prediction each time new transitions have been found.

3.1.2 Structural Perturbation Method

Next the structural perturbation method (SPM) introduced by Lü et al. [2015] is analyzed. This method is based on the eigenvectors of the adjacency matrix, which capture the many important features of spectroscopic networks as shown in section 2.5. The eigenvectors are modified by the same calculation used for perturbation theory in quantum mechanical systems. It is important to note that no quantum mechanical perturbation calculation of the spectrum is performed. Instead, a random dropout (in addition to the one used for the dropout method) is used to do a first order perturbation theory calculation of the eigenvalues.

The algorithm works as follows:

1. Split the network into two sets: 10 % of the links are seen as a perturbation, the other 90 % are the remaining links.

2. This corresponds to a splitting of the adjacency matrix into a perturbation $\Delta\mathbf{A}$ and a remaining part $\mathbf{A}^R$.

3. Diagonalize $\mathbf{A}^R$

4. Do a first order perturbation theory calculation with $\Delta\mathbf{A}$ as a perturbation of $\mathbf{A}^R$ to find corrections $\Delta\lambda_i$ to the eigenvalues λ_i of $\mathbf{A}^R$
5. Calculate $\tilde{\mathbf{A}} = \sum_i (\lambda_i + \Delta\lambda_i)\vec{v}_i\vec{v}_i^T$
6. Repeat 1 to 5 many times and take the mean of $\tilde{A}$ over all repetitions
7. The entries in the mean of $\tilde{A}$ assign a ranking to all possible links

The motivation behind this algorithm is: The eigenvectors of the adjacency matrix capture many important structural features of the network, thus they should only undergo minor changes due to the missing links in the network – unless the missing links completely change the network structure, in which case they are hard to predict.

Again the prediction quality can be visualized using ROC curves (figures 3.2 to 3.4). For helium this prediction is almost perfect: The true positive rate (TPR) rises to about 40 % with a false positive rate (FPR) of almost 0. Further, the false positive rate increases only to about 3 % for a true positive rate above 90 %. This means that 40 % of the true links are found again almost without errors, and 3 % of the false transitions are found within 90 % of the true transitions.

For iron still the TPR still increases to roughly 40 % immediately, but the FPR increases to 5 % for a TPR of 90 %. After that the curve stagnates at about 95 % TPR. This means that the method performs poorly at recovering the last 5 % of true links: A random choice is always given by a diagonal from the current position to (1, 1). Starting where the curve hits 95 % TPR the ROC curve is lower than this diagonal, thus the prediction performs worse than random here. This indicates that some transitions are hard to find using the SPM.

For thorium the first jump of the TPR goes only to about 20 %. of the links can be found with very little error, and afterward the curves flattens out earlier than for iron: The FPR rises above 10 % for a TPR of about 90 %. At some point the curve becomes completely flat, indicating that there is no true link found between 30 % and 85 % FPR.

3.1.3 Evaluation

Both methods, the eigenvector based SPM and the statistics based NSBM work well for link prediction. This is first of all a further indicator that the eigenvectors and the communities found by the NSBM are good characterizations of the graph, which confirms the results of section 2.5. While the NSBM works best for iron and thorium,

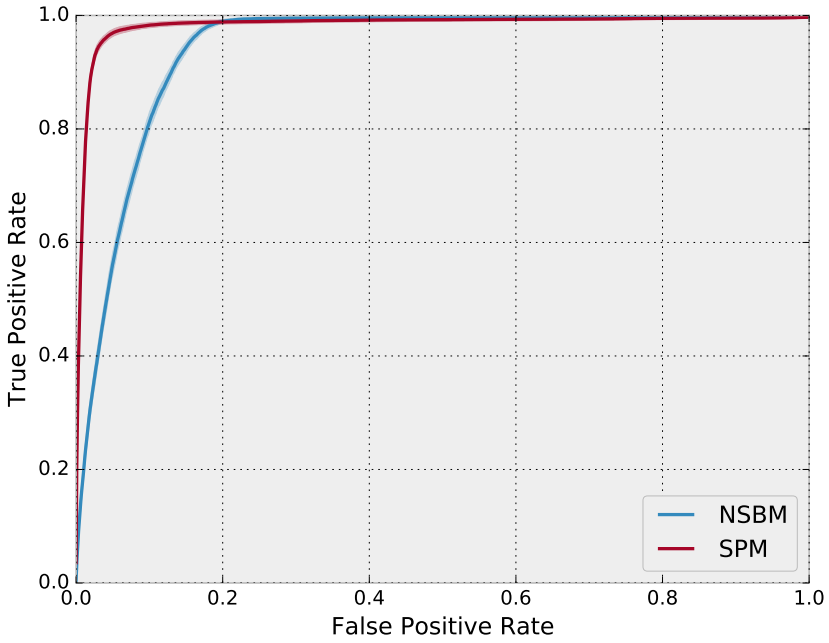


Figure 3.2: Helium ROC curves for NSBM and SPM with a 10 % dropout.

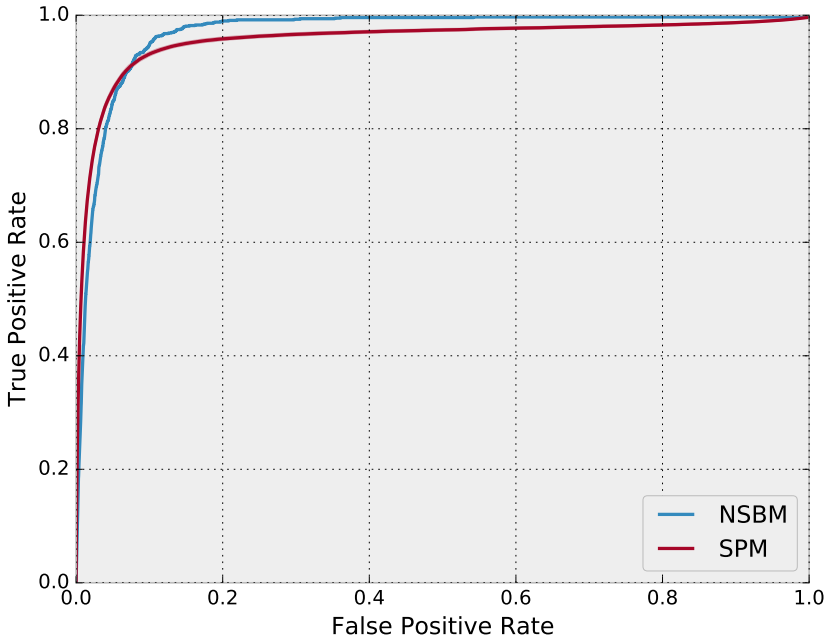


Figure 3.3: Iron ROC curves for NSBM and SPM with a 10 % dropout.

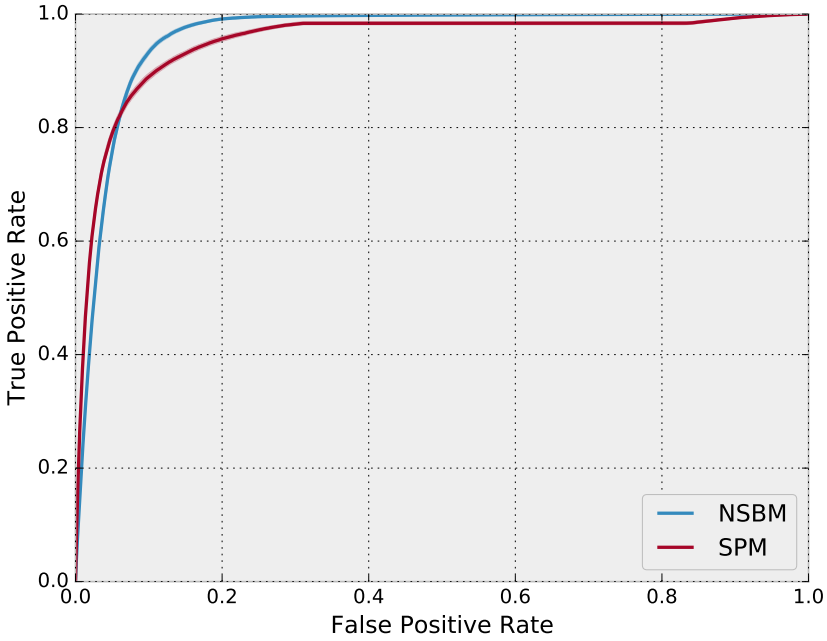


Figure 3.4: Thorium II ROC curves for NSBM and SPM with a 10 % dropout.

the SPM works best for helium. This indicates the strength and weaknesses of the different methods: The NSBM needs a lot of data to find a statistically significant community structure. If this data is available, like for iron and thorium, it does very well at finding the structure and using it for predictions. Since the SPM works with eigenvectors, statistical significance does not matter here. Instead, it is important that the underlying structure can be captured well by eigenvectors. For the ladder-like LJ structure of helium (sometimes including S) this has been shown to be the case. For iron and thorium the structure becomes more complex and is thus less accessible with the SPM.

This does not mean the SPM works poorly for iron and thorium II. Even for iron and thorium II the SPM performs better than the NSBM in predicting the most likely links, showing that an important part of the structure is captured by the eigenvectors of the network. The correspondence of eigenvectors and NSBM to atomic structure has also been analyzed in section 2.5 with respect to predicting quantum mechanical properties. It was found that on the one hand for helium quantum numbers can be predicted almost entirely using eigenvectors, whereas the NSBM performs slightly worse. On the other hand the NSBM operates better for the atoms iron and thorium II. For these atoms, the eigenvectors perform poorly at

predicting the configurations. Nevertheless, the eigenvectors function well for link prediction for all atoms, indicating that the configurations given are not relevant for the transitions.

For iron and thorium the strengths of the two methods are complementary: The NSBM finds essentially all of the actual network with few errors, whereas the SPM is not able to recover the last percents of links. This feature is of interest if there is a new line seen and it is unclear which element it belongs too (for example because it is in some stellar spectrum). In this case the NSBM can be used to decide whether two states with the correct energy difference can have a transition between them. In contrast, if the SPM assigns a low likelihood, the transition might be part of the last percents only recovered in the end.

The strength of the SPM is the very steep slope in the beginning. Even though the NSBM has a steep slope as well, the SPM clearly does better at the start for all atoms analyzed. As discussed above, this means most of the links which get a high likelihood assigned to them are part of the true network. This is particularly important if new links shall be proposed for a measurement. Here, only the links with the highest likelihood will be taken into account, so that only the slope at the start of the curve is important.

3.1.4 Conclusion

So far only a random dropout of 10 % links was considered. This leaves to questions open: "What happens for more than 10 % missing links?" and "What happens for a non-random dropout?" Both questions are important since there is no insurance the experimental data bias can be modeled by 10 % random dropout. In his thesis, Julian Heiss answers the first question and shows that the methods still work well for a larger dropout fraction.

It should now be checked if the methods still work well in a more realistic scenario. Which transitions are measured and which are not is not decided randomly, but strongly dependent on the wavelengths and intensities of the different lines. Different measurement devices are only able to measure certain wavelength ranges and have an intensity threshold. As the wavelength range broadens and the intensity threshold decreases, new transitions become accessible and thus the knowledge about the spectrum advances. It is not obvious whether a random dropout, which treats all links equally, is sufficient to model this. To take the bias into account it would be best to compare the methods to new measurements or use historic datasets and try

to predict the now known lines starting from there. Then these results could be more confidently extrapolated to new measurements.

This measurement bias also creates another problem: For complex atoms the ground truth is not completely known. Many transitions which are counted as false positives might actually appear in nature and are just not measured yet. This can influence the result in two ways. On the one hand, if these transitions are considered likely by the algorithm, the actual result would be better than the one we calculated. On the other hand, if the false positives are considered unlikely the quality of the result overestimates the method quality. It could be visualized on the ROC curve as either making the beginning steeper (first scenario), or making the end steeper (second scenario). To solve this problem the ground truth needs to be known completely. This is the case for helium, which is why helium is a good benchmarking system. The structure of helium is much simpler than that of iron or thorium, which can also effect the prediction accuracy, and there is no easy way of getting access to the ground truth for complex atoms. Here the only way to improve on the ground truth is to check for all lines predicted and so that can be confirmed or falsified.

Another natural extension of the link prediction that could also help tackle the above described issue is the prediction of weights. As discussed in section 2.2 the dipole matrix element is a good weight choice. Predicting the weights of transitions, it could be checked whether the transition should have been measured: The wavelength is known from the energies and the intensity can be calculated from dipole matrix element and wavelength. Implementing weight prediction is generally a hard problem, but some steps have already been taken by Peixoto [2018a]. This would also allow us to take into account that all transitions are possible, many of them just have a vanishing weight and should consequently not be considered. For a binary choice: strong transitions are considered, weak ones are not, the line between strong enough to be considered and to weak is arbitrary. For experimental networks like iron and thorium this question is not as important, because mostly strong transitions have been measured. For the calculated helium network the question where to draw the lines is less clear to answer: Should all singlet-triplet transitions be considered or not?

Contrary to most first principle quantum mechanical calculations the NSBM and SPM do not give a clear “Yes” or “No” for any transition, but only a more or less likely. Therefore this prediction is a lot harder to evaluate quantitatively. Usually

this is done by the ROC curve, but there is no clear cut on what is a good ROC curve and what is a bad one. The NSBM assigns actual probabilities to the transitions (without absolute normalization), hence a quantitative analysis and comparison of these probabilities to a measured ground truth might be possible. The SPM only assigns a score to each link, which cannot be translated into a probability, i.e. twice as high does not mean twice as likely. This means that there is no quantitative evaluation that goes beyond the ROC curve and thus the interpretation can be subjective.

Another step towards better link prediction would be to develop a new method specifically tailored to spectroscopic networks. So far, only general methods were used, which do not consider certain features of our network such as bipartivity. Furthermore our knowledge of quantum mechanics could be introduced to make a joint prediction about quantum mechanical properties and new links. The NSBM already goes into this direction, but developing a method particularly tuned to this type of network might offer new insights into quantum mechanics of complex atomic systems. A first step could be to combine the NSBM and the SPM, to profit from both' strengths. For example one could check which transitions are predicted by both models and which only by one of them.

All in all both methods – NSBM and SPM – work very well for predicting links, but there is still some space for improvement. This means on the one hand we have a confirmation that these methods and their underlying features – inferred communities and eigenvectors – are correlated with the underlying physical laws describing these transitions. On the other hand these method can be used to guide future measurements, greatly reducing the experimental effort.

3.2 Node Prediction

In the previous section we showed how link prediction can be used to predict new transitions. In this section we analyze the problem of predicting new states by means of the network. This problem is hard to tackle using standard approaches in physics, as finding the eigenvalues and eigenstates of a Hamiltonian with many strongly interacting particles is an open problem. This problem was tackled by Safronova et al. [2014] for thorium III, which has only two outer electrons, but for thorium II only approximate values for the energies and no values for the transition strengths were given. Therefore new approaches are needed, and in this section we propose a new approach based on network science.

Finding new states is of practical interest: New states can give a deeper insight into the structure of the atom, and they might have interesting electronic properties, making them relevant for further physical analysis. An example for such states is the ongoing search for new high energy states in thorium ions. These states could couple to the low lying nuclear excitation. This excitation is very narrow and would thus be a candidate for a nuclear atomic clock, which would be better than the current atomic clocks working with electronic transitions (Peik and Tamm [2003], Herrera-Sancho et al. [2013]).

Unfortunately there is no simple way to extract all properties of a new state from the network. Thus, we will merely address the subproblem: "Which transitions do the new states have?" Answering this question will not lead to energies of the new states; finding a way to predict those lies beyond the scope of this thesis. Translated into network theory we ask: "Can the links of new nodes be predicted?"

3.2.1 Introduction

Because direct information on the connectivity of the new nodes is missing, it is hard to make predictions about their links. This problem has been analyzed in the literature as the cold start problem of recommender systems by Son [2016], Lika et al. [2014] and Wei et al. [2017]. There, the problem is given as follows: Consider the network of products and users of an online shopping portal like Amazon. The nodes are the users and the products, and a link is added between the user u and the product p if u buys p. Now consider a new user joining the shopping portal or a new product being offered. It needs to be examined which products can be recommended to the new user or to whom a new product can be recommended. These questions are

known as the new user cold start problem and the new product cold start problem respectively. Usually recommendations are based on link prediction; but as was seen in the last section, standard link prediction needs links for the nodes. A further distinction is made into the complete cold start problem and the incomplete cold start problem, where for the latter a few links are already known for the new nodes. Especially for the complete cold start problem little literature is available.

In a recent review Son [2016] compared different approaches to tackle the cold start problem, focusing on the incomplete cold start problem. Approaches of using only the available links for prediction, approaches using meta data for prediction and combinations of both have been compared, and the methods which worked best were the ones only considering the available links. This shows that taking away these few links, as is the case in our scenario, makes the problem significantly harder.

The usual approach to the complete cold start problem is to use the meta data to find a group of similar nodes. Then the links of the new node are given by the links of the other nodes in the group. Predicting which users or products are similar can be done for example by neural networks (see Wei et al. [2017]) or by traditional clustering. A relevant difference between recommender systems and atoms is the different size of the networks. As usually millions of links and tens of thousands of nodes are in one of the networks considered for recommender systems, the algorithm that is used has to be as fast as possible and should make the best use of the enormous data amount. This usually comes at the cost of complexity. In contrast, the networks we deal with contain only a few hundred nodes and a few thousand links, which enables us to use the global network structure with eigenvectors and the NSBM. Methods based on neural networks in contrast need much more data and are thus not fit to be used on spectroscopic networks.

There are further approaches to node prediction by Kim and Leskovec [2011], Eyal et al. [2013] and Hric et al. [2016]. Kim and Leskovec [2011] assume that the number of nodes in the network is a power of 2, Eyal et al. [2013] assume that the links towards the missing nodes are known. Both of these assumptions are invalid for spectroscopic networks. Hric et al. [2016] assumes the further use of discrete meta data. This is a reasonable assumption for spectroscopic networks, but there is no simple implementation available of his results and redeveloping the necessary code is beyond the scope of this thesis.

Thus we develop a new approach with assumptions based on the physical properties of spectroscopic networks. The assumptions will be similar to those made by

using meta data for recommender systems, as it is reasonable to assume some prior knowledge of the nodes to predict. As spectroscopic networks typically only have a few hundred nodes, we can use methods based on the global structure like the NSBM and eigenvectors. These have also proven to work well in previous chapters.

3.2.2 Assumptions

In order to predict the links, we will assume that the NSBM communities of the new nodes are known. For thorium II in sections 2.3 and 2.4 it was indicated that the community can be determined if J and the configuration are known. This can be seen in the figures 2.7 and 2.13: Knowing the parity (part of the configuration) and J results in the knowledge of the level 1 community, and knowing the configuration in addition is usually enough to also guess the correct level 0 community.

Now we will argue why configuration and J pairs should be known. First of all are the individual accessible orbitals for the single electrons well known, because these are just all unoccupied orbitals. For example for thorium II the electrons can access the orbitals $7s$, $7p$, $6d$, $5f$ and higher. The electrons can now occupy any combination of those orbitals, the only constriction is: a maximum of two electrons per s-orbital, six per p-orbital and so on. One such configuration would be: $7s7p6d$. With simple combinatorics, for each configuration the different possibilities for J can be found: $J \in \{|L-S|, \ldots, L+S\}$. For $7s7p6d$ this leads to: $S \in \{1/2, 3/2\}$, $L \in \{1, 2, 3\}$ and $J \in \{1/2, \ldots, 9/2\}$. Thus the different possibilities for J and configuration are known.

For a certain combination of configuration and J the number of states can be calculated. This was done by Zalubas and Corliss [1974]; more recent calculations were done by Safronova et al. [2014]. This would be the number of states with pure configurations, which do not appear in nature, but it also gives the dimensionality of a respective subspace in the Hilbert space. As the dimension of a Hilbert space is independent of its basis, the number of states with a certain pair of dominant configuration and J can be guessed to be similar to the dimension of the respective subspace itself. Combining this with the community knowledge, the number of nodes in each community can be guessed. This number will not be perfect, but it can give a clear indication for which community new nodes are needed – for example if only half of the guessed nodes are known. Thus we can assume that for each added node the community of this node is known.

3.2.3 Group Method

A straight forward approach to use the community information in order to predict transitions for the new states is: Use these communities for NSBM link prediction. The NSBM link prediction is described in section 3.1 and appendix B. It can be imagined as follows: A new node in community A should be connected to the same communities as the other nodes in community A. In addition modifications according to node degrees and changes in the network due to the additional links are taken into account. Instead of sampling different community structures as in section 3.1, the communities are now hard coded based on the assumption made.

In order to evaluate the results, a random dropout is used and the result quality is visualized by a ROC curve. The following steps are taken:

1. The community structure of the *ground truth network* (the network before dropout) is inferred using the NSBM.

2. 10% of nodes are randomly chosen and all their edges are removed. These are now considered as the *new nodes*.

3. For every potential link involving at least one new node a likelihood is calculated using the NSBM.

4. The potential links are sorted by likelihood and compared to the ground truth network.

5. The results are visualized using a ROC curve as described in section 3.1.

These ROC curves are shown in the figures 3.5 (helium), 3.6 (iron) and 3.7 (thorium). The results found show that this prediction works well: In helium, the true positive rate (TPR) rises to about 85 % for an false positive rate (FPR) of 5 %. For iron and thorium the results are slightly worse: For an FPR of 10 %, the TPR is at 80 % (iron) and 85 % (thorium). At these points the slope of the ROC curve starts to decrease, so that the ranking beyond this point is only slightly better than random. This decrease in prediction accuracy founded in special features of the individual nodes, which might be different from other nodes in the group and thus cannot be predicted by the NSBM.

3.2.4 Eigenvector Method

Next, we will use eigenvectors to predict new nodes. They were already used successfully to predict quantum numbers (section 2.5) and transitions (3.1). This indicates that they are good features to describe the structure of spectroscopic networks. In order to decide how eigenvectors can be used to predict new nodes, it is first analyzed how the spectrum changes when nodes are randomly removed. This can then be flipped around to add missing nodes again. This analysis can be found in appendix C and yields the following results for the adjacency matrix:

1. The eigenvalues scale with network size.
2. The components of eigenvectors with large absolute eigenvalues are not changed (up to normalization).

Hence, the following features are required of the node prediction algorithm: The new eigenvalues should be similar to the old ones and the components of the large eigenvectors (eigenvectors who's absolute eigenvalues are large) of the known nodes should not change by much.

In order to calculate the matrix from the eigenvalues and eigenvectors, the following calculation will be done:

$$\mathbf{A} = \sum_i \lambda_i \vec{v}_i \vec{v}_i^T \tag{3.1}$$

Here the λ_i and $\vec{v}_i$ are the eigenvalues and eigenvectors respectively. This shows that changes of small eigenvectors have only a small influence on the adjacency matrix: For $\lambda_i = 0$, $\vec{v}_i$ has no influence on the spectrum at all. Thus, only a minor error will be made if the small eigenvectors are left roughly constant as well.

To find the new adjacency matrix, the following additional information is needed about its spectrum: "What are the components of the eigenvectors for the new nodes?" and "What are the additional eigenvalues and eigenvectors?"

To answer the first question, consider the following feature: the components of the (large) eigenvectors in the same group are similar. This has been used in section 2.5 to predict the quantum numbers of nodes. For the problem of node prediction this means that the new components that need to be assigned to the eigenvectors for the new nodes should resemble those of the other nodes in the same group.

To answer the second question, we can assume that the new eigenvalues are small. As the large eigenvalues scale with network size, the missing eigenvalues are not

among those and hence can only be the small ones. This means they can be approximately set to zero and omitted from the analysis.

We thus propose the following algorithm, which has the desired properties:

1. Diagonalize the adjacency matrix.
2. For each eigenvector of the adjacency matrix, add as many new components as there are new nodes.
3. Each new component has its value assigned as follows:
   a) Find the community of the node belonging to this component.
   b) Take the average of all other nodes in this community. This is the new component.
4. After the new components are assigned to the eigenvectors we can transform this back to a matrix $\tilde{\mathbf{A}}$ as:

$$\tilde{\mathbf{A}} = \sum_i \lambda_i \tilde{\vec{v}}_i \tilde{\vec{v}}_i^T \tag{3.2}$$

   λ_i are the old eigenvectors and $\tilde{\vec{v}}_i$ are the modified eigenvectors
5. $\tilde{\mathbf{A}}$ is now an approximation to the new adjacency matrix, and its entries give a score to each link.

In appendix C.2 we prove that this fulfills the condition of only slightly modifying the eigenvectors. The eigenvalue conditions on the other hand cannot be proven easily.

The algorithm can be applied to the Laplace matrix as well, since it does not care which matrix I give it to modify – as long as it is diagonalizable and each row corresponds to one node. The Laplace matrix is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$ with $\mathbf{D}$ being the matrix with the node degrees on its diagonal. The link scores are then given by the respective components of $-\tilde{\mathbf{L}}$.

The results for both adjacency and Laplace matrix are shown in figures 3.5, 3.6 and 3.7. It can be immediately seen that adjacency and Laplace yield the same result. The reason for this is not known. The eigenvalue spectra of both matrices encode very different properties of the graph, so it is also not clear why this should be the case. In the method described, the eigenvectors and not the eigenvalues were changed, and those might have an easy relation in the case of adjacency and Laplace.

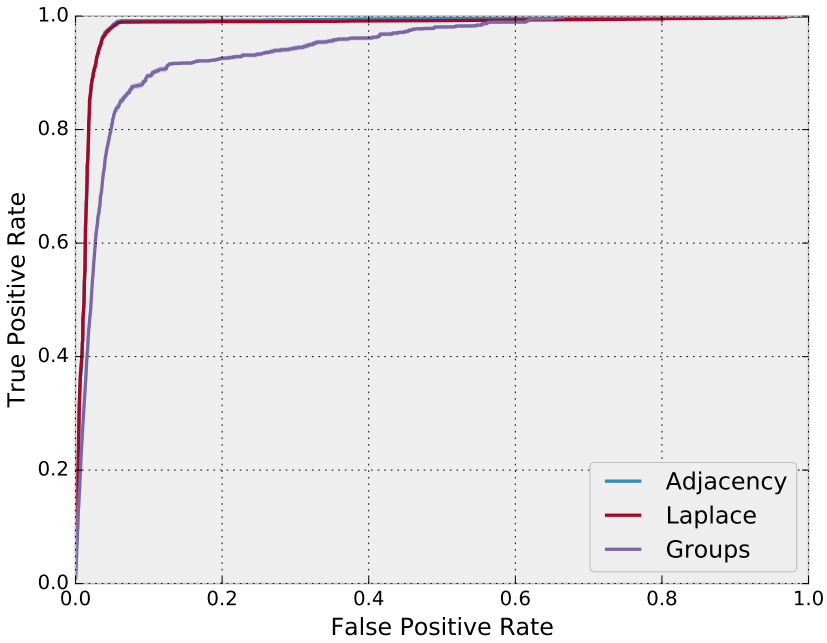


Figure 3.5: Helium ROC curves for different node prediction methods with a 10 % dropout.

Again, the algorithm works by far the best on helium: For a FPR of 5 %, almost all true links are recovered (99 % TPR). This also confirms that the helium network is well described by its eigenvectors. For iron these values go down to 90 % TPR at 5 % FPR and for thorium 95 % TPR at 10 % FPR. At these values the ROC curves have a kink and for links with low scores the prediction is roughly random.

For helium and iron the ROC curve almost follows the y-axis in the beginning. As discussed in the previous section, this means that the high scoring predictions are very accurate. This feature is very important, as these predictions can then be used as guesses for the new links and these guesses are mostly correct. For thorium the ROC curve clearly has a finite slope even in the beginning, but as this slope is large, the prediction is still good.

3.2.5 Comparison

We have shown that knowing the community of a state is sufficient to predict its transitions. We have found that using a naive approach, i.e. applying link predictions with known communities already yields good results. The prediction accuracy can

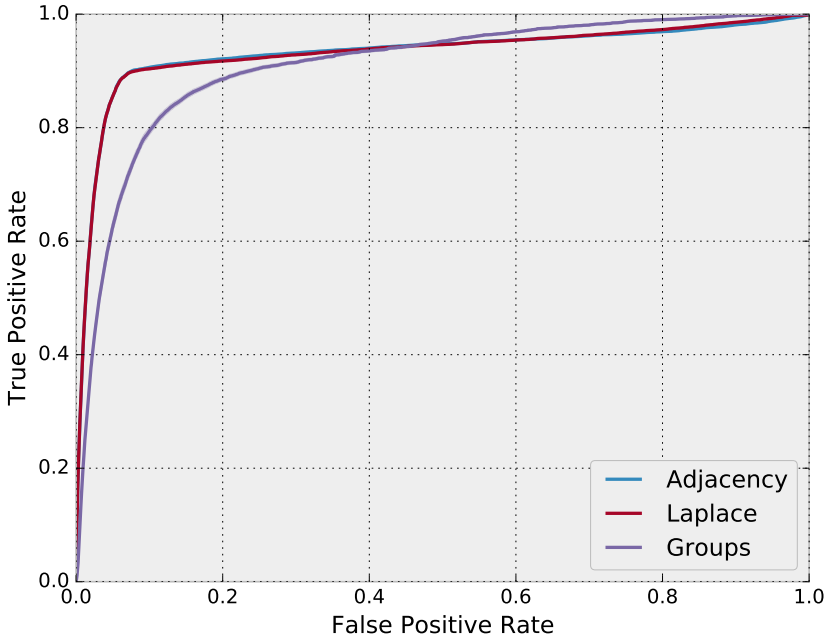


Figure 3.6: Iron ROC curves for different node prediction methods with a 10 % dropout.

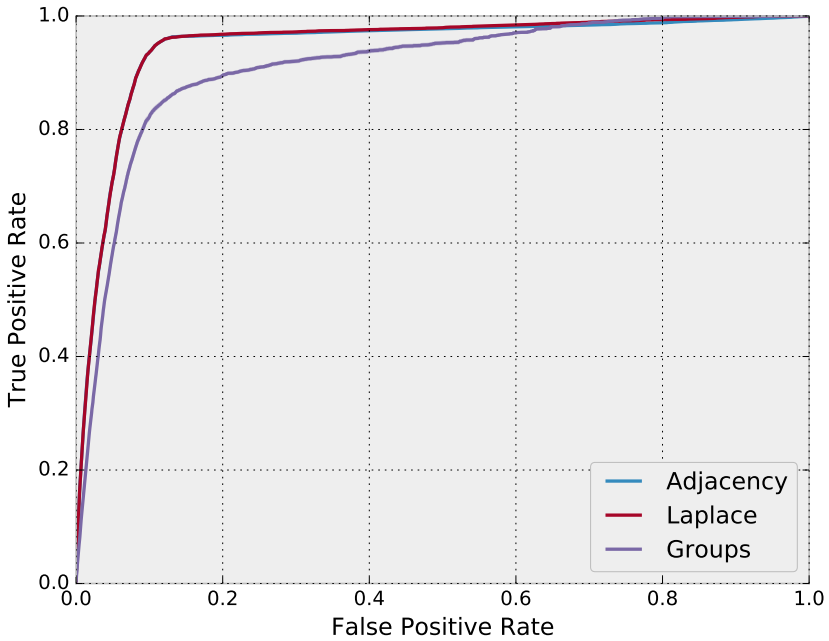


Figure 3.7: Thorium II ROC curves for different node prediction methods with a 10 % dropout.

be significantly increased if, in addition to the community structure, eigenvectors are taken into account. Here, it does not matter whether the eigenvectors of the adjacency matrix or the ones of the Laplace matrix are considered.

For both methods – the NSBM and the eigenvector method – the prediction separates into two parts: a steep increase in the beginning and a flat part in the end. For the NSBM method, the transition between the two parts is slow, whereas for the eigenvectors there is a sudden transition leading to a kink in the ROC curve. Using the eigenvector method, for helium essentially all true links are recovered at 5 % FPR, whereas for the NSBM method these are only 80 % to 90 %. For Iron with the eigenvector method $\sim 90\,\%$ of true links are recovered at $\sim 5\,\%$ FPR, for thorium II the ROC curve rises to $\sim 95\,\%$ TPR at $\sim 10\,\%$ FPR. In all three cases the group method performs significantly worse.

Thus we find that using the eigenvectors improves the method considerably: The initial slope of the ROC curve is steeper and more links are recovered before the ROC curve flattens. These are two very important features, as they indicate that a higher fraction of the proposed links (links with very high score) are part of the actual network and a bigger part of the network can be recovered before mostly false links are predicted.

For all three atoms examined – most notably for iron – the ROC curves of the group method and the eigenvector methods intersect. This indicates that if the $X\,\%$ modt likely predictions are considered (for iron $X > 50$, for helium and thorium $X > 80$), the NSBM finds more true links than the eigenvector method; more precisely at an FPR larger than X, but since almost all potential links are false this is a minor difference. There are few cases for which links at such a high FPR need to be considered, so this is an interesting feature with no immediate relevance. A similar feature was found for link prediction in the previous section for the comparison of SPM and NSBM, but at much lower FPR.

3.2.6 Conclusion

The error sources for node prediction are similar to those for link prediction. The main error source is probably the random data dropout. Dropping out nodes at random is particularly unrealistic, as usually nodes with very high energies or J or other extreme properties are unknown, whereas nodes close to the ground state are well known. With a random dropout nothing prevents the ground state from being dropped, creating a totally unrealistic scenario from the physics point of view. As

in the case of link prediction this could be solved by either comparing to future experiments or use historical data and compare predictions to today's dataset.

On top of that the assumption of knowing the node's community is very strong. First of all it was shown in section 2.3 that the correlation of the communities with the quantum numbers is not perfect, so there will be errors if the communities are assigned based on their quantum numbers. It remains to be checked if this has positive or negative influences, as it means that nodes are assigned to communities closer to their physics, which could also improve the prediction. As the configurations are only partly known and it is not known which combinations of configurations are in the new states, the exact number of nodes per community is hard to guess, but a rough estimate should be enough for good results.

Related to this error is the error made by assigning communities before the dropout. In a realistic scenario the additional data is not available though, so the communities have to be assigned without the additional nodes. This method was necessary to model the group knowledge, but a better method would be to run the NSBM after node dropout and then add the nodes to the correct communities based on their quantum numbers.

In order to guide new measurements with the predictions, it would be necessary to also assign energies to the new states. Unfortunately, we know no promising approach to determine the energies from the network. A basic idea could be to use the energy-community-relation from section 2.4 or to guess energies from the configurations. But this naive approach leads to a broad range of energies, whereas a very exact determination would be necessary to guide measurements. Hence, it might be best further develop quantum mechanical calculations like configuration interaction (see Safronova et al. [2014] for thorium) to determine the energies and configurations of the states and use the network to determine which transitions are allowed.

In this chapter we have proposed a way to tackle the problem of node prediction assuming the communities of the new nodes are known. The method performing best in finding most of the true links quickly and in making few errors in the beginning is using the eigenvectors of either the adjacency matrix or the Laplace matrix.

The introduced method allows us to predict the transitions of so far unmeasured states. Speaking in another context: We are able to predict for which pairs of states the dipole matrix elements are non-vanishing. This method could be used as a tool in a bigger framework that enables theoretical predictions about eigenstates of

complex atomic Hamiltonians and their transitions.

4 Conclusion

In this thesis, we have atomic spectra with methods from network theory, which enabled us to find quantum mechanical properties and make predictions about the network. Established methods as well as methods developed for specific tasks were successful and showed the potential of this novel approach.

In chapter 2, spectroscopic networks were introduced and their structural features were related to quantum mechanics. First, the parity was related to bipartivity and an algorithm was developed to find the sets of states of different parity and the transitions between them. This algorithm was proven to work under reasonable conditions.

Community detection was employed on the network in order to find further groups of similar states. We discovered that with only a few exceptions, all states that are in the same community have similar quantum mechanical properties. For helium, this means that they have equal orbital angular momentum L and total angular momentum J; for iron and thorium they have equal parity and J. These are also the quantum numbers that are important for the selection rules and thus should be important for the structure of the network.

For iron and thorium, a smaller splitting of the communities indicates additional structure, as for one set of parity and J multiple communities were found. For thorium, this additional structure corresponds to the formation of clusters in the energy domain: each community forms either one or two clusters of states with similar energies. The community structure also appeared to have some correlation with the configurations that can be used to describe the respective states. For communities that separate into two clusters, the states in the high energy cluster have dominant configurations including a $7p$-orbital, whereas the low energy cluster contains no $7p$-orbitals.

These results can be used to predict the parity and total angular momentum for single unknown states in iron or thorium. The orbital angular momentum of helium can be predicted as well. For helium, eigenvectors seems to be more efficient for this prediction, whereas for iron and thorium the prediction showed better results when

the communities from above are used.

In chapter 3, different methods were used to make direct predictions about the spectrum. First, new transitions between known states were predicted. The methods find a ranking of all possible links agnostic of quantum mechanics based on the network structure. Here two methods were analyzed: the NSBM based on the aforementioned community detection and the SPM based on eigenvectors. These methods were benchmarked with a random dropout of known transitions and it was shown that both methods work well. For helium, the SPM is more efficient, whereas for iron and thorium both methods have different strength: The SPM initially makes more correct predictions, the NSBM is faster at recovering all deleted transitions.

In addition, new methods were developed to predict transitions to new states. Here, a prior knowledge about the communities of the nodes was assumed. It was shown that the eigenvectors provide good predictions of the new states. These methods were again benchmarked with a random dropout and it was shown that many of the deleted links can be recovered with high accuracy. Less than 10 % of the links to the deleted nodes were difficult to rediscover.

For the methods analyzed in this thesis much spectral data is needed. As network science is a data based analysis, it leads to better results if more data is available. This can be seen, for example, in the resolution limit of the NSBM: As the minimum size of the group scales logarithmically in the number of nodes of the network, the potential number of communities found grows with the network size. If only little data is available for one specific feature, it cannot be detected, and if little data is available overall only few features can be detected. As there is much data known for the interesting atoms, the methods show good results on these atoms.

It is not clear how the methods presented here interact with the measurement induced bias that only certain wavelengths can be measured. For example, the additional structure found with community detection for iron and thorium might also be caused by this data choice.

For link and node prediction, only random dropouts were considered to benchmark the model, but actual new transitions will be discovered when new tools become available that allow measurements of new wavelengths. Similarly, new states are likely to be discovered at high energies. This means that the link prediction is no benchmark against systematic data selection.

In order to analyze the limitations of data selection, the data could be artificially modified by introducing a new minimum and maximum wavelength. The effects of

this data modification could then be analyzed.

To test the performance of link prediction and node prediction in a more realistic scenario, historic data could be used as the basis of prediction. Even better would be to use all the now available data for prediction and compare this prediction to future measurements.

So far, methods that are to a large extent agnostic of the underlying principles of the network were used. The NSBM only assumes the existence of communities and the SPM only assumes that eigenvectors are a good feature to describe the network. To improve on these methods, it might be possible to include specific features of the network into the prediction. One such example could be to explicitly use the bipartite structure of the dipole transitions to develop better community detection.

Another step to improve the model would be the inclusion of weights that include specific atomic properties. We already proposed to use $A_{ik} * \lambda_{ik}^3$ as a weight. It would be interesting to generalize the link prediction to a weight prediction, which assigns guessed weights to the predicted links. Furthermore, the weights could also be incorporated in the community detection in order to improve its results.

It might be possible to find additional features in the energies and use these to add energies to the states predicted by node prediction. This could either be done by introducing node weights, or by calculations based on atomic physics.

Spectroscopic networks are the only network, for which a physical ground truth is available. Since this means the ground truth is completely known and can be related to the laws of nature, they are an ideal benchmark system for network methods. This is especially interesting for methods analyzing bipartite networks, where it could take the role that the infamous karate club of Zachary [1977] has for community detection. This is also true for link prediction and weight prediction algorithms, as all links and weights are known.

If the additional groups found for iron and thorium II turn out to be based on atomic properties, they could lead to a deeper understanding of the atomic structure. The communities might turn out to be states with similar physical properties, which would help to find new weak symmetries in the atoms. These could then in turn be used to enable better, approximate first principle calculations to analyze the quantum mechanical structure of the atom. Such calculations would lead to a better understanding of the atom as well as its chemical and optical properties.

The methods used here are not bound to atomic spectra, but can in principle be applied to all discrete spectra. Prime examples of this would be molecular and

nuclear spectra. Much data is available for these spectra and they have a similar fundamental structure and parity albeit also many unique features.

All in all, we have introduced network science as a new framework with which to analyze atomic spectra. This framework enables us to apply methods from network science onto nuclear spectra. With this method, we were able to identify similar states and relate these to well known quantum numbers without introducing them as physical knowledge beforehand. We were also able to show that network methods can be used to predict transitions, so that the methods can be tested against measurements. This prediction can further be used to guide the search for for new transitions. Furthermore, ideas how the network could be used to predict new states were presented.

Appendix

A Energies

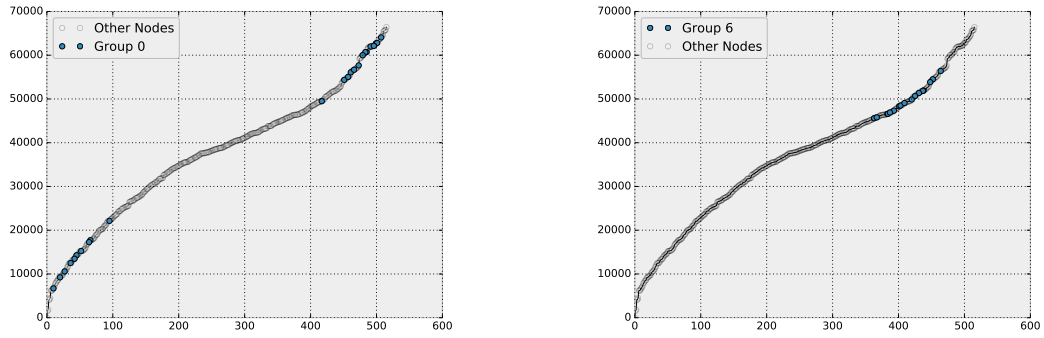
In this appendix some further information on the energies of the states is given. This further information can help to decide whether the structure found in section 2.4 is the result of measurement artifacts or fundamental physics.

A.1 Comparing one Group to the Entire Spectrum

In order to check, whether this pattern connects in a simple way to the total energy spectrum, we will plot how the groups are embedded in there. This can be seen in figure A.1. Here, we observe that the nodes of these two groups are clustered. For group 0 we find two clusters and for group 6 one cluster, but there is no clear structure within these clusters.

A.2 Results for Other Elements

The transition energies of other elements are considered in this section, to test whether the relation to the communities found in thorium II was caused by the transition energies or not. Some of them like thorium I show the same energy



(a) Group 0 (b) Group 6

Figure A.1: The groups 0 and 6 embedded in the complete Th^+ spectrum.

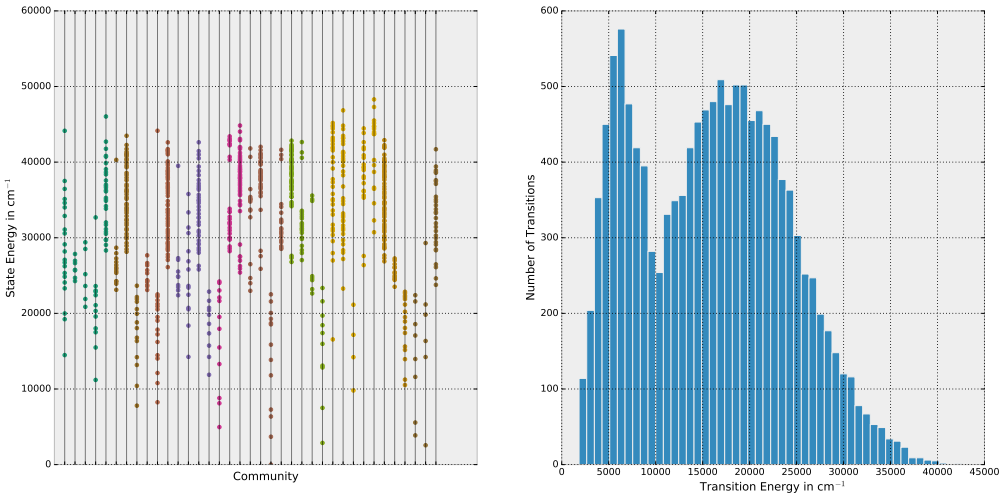


Figure A.2: Energies and Communities in Th I

clusters as thorium II, others like iron I do not have this structure. There always seems to be some structure in the energy, but it seems to be more or less prominent for different atoms. A transition energy structure consisting of two blocks like the one of thorium II is not found for the other atoms, but peaks at certain energies can still be found (for example 25 000 cm^{-1} to 50 000 cm^{-1} for manganese II).

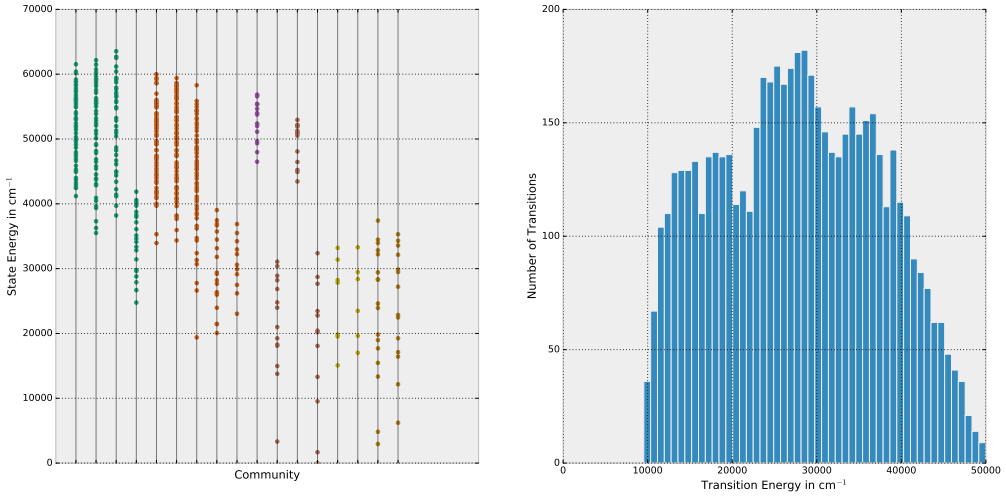


Figure A.3: Energies and Communities in W I

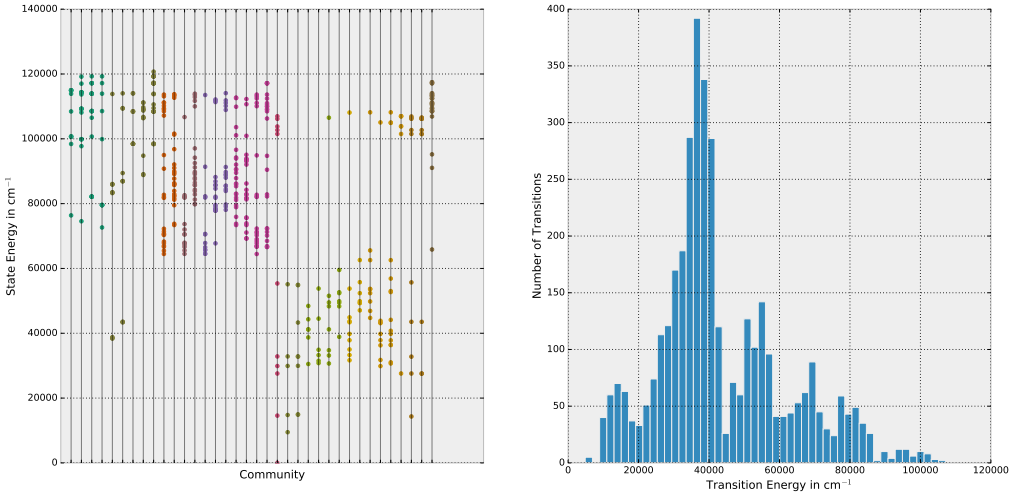


Figure A.4: Energies and Communities in Mn II

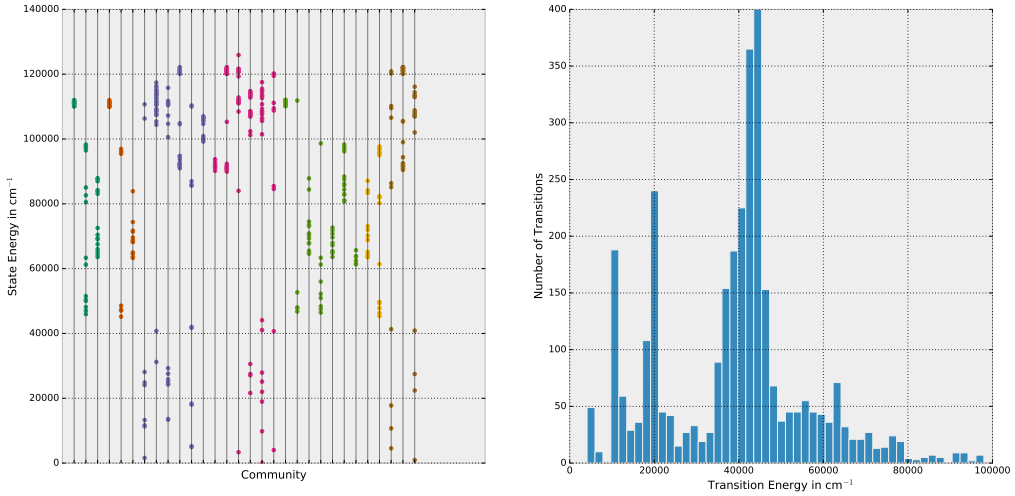


Figure A.5: Energies and Communities in Co II

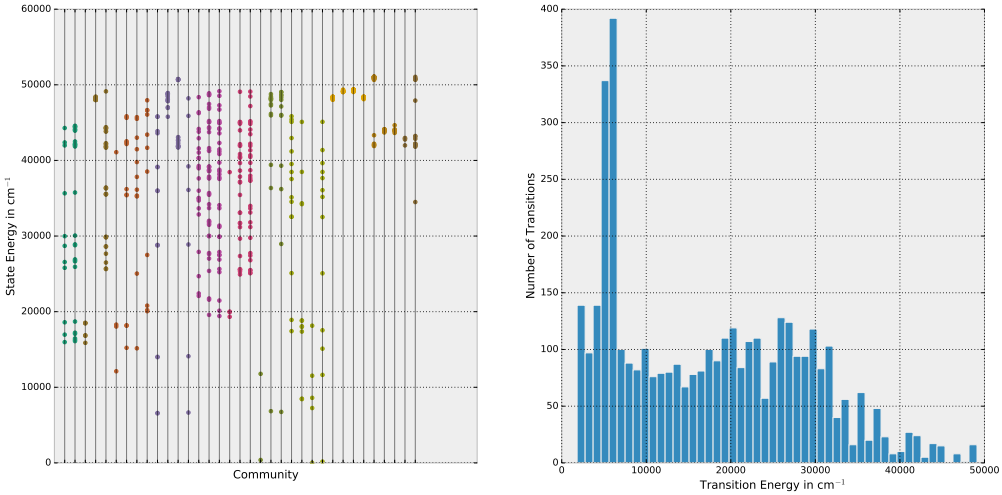


Figure A.6: Energies and Communities in Ti I

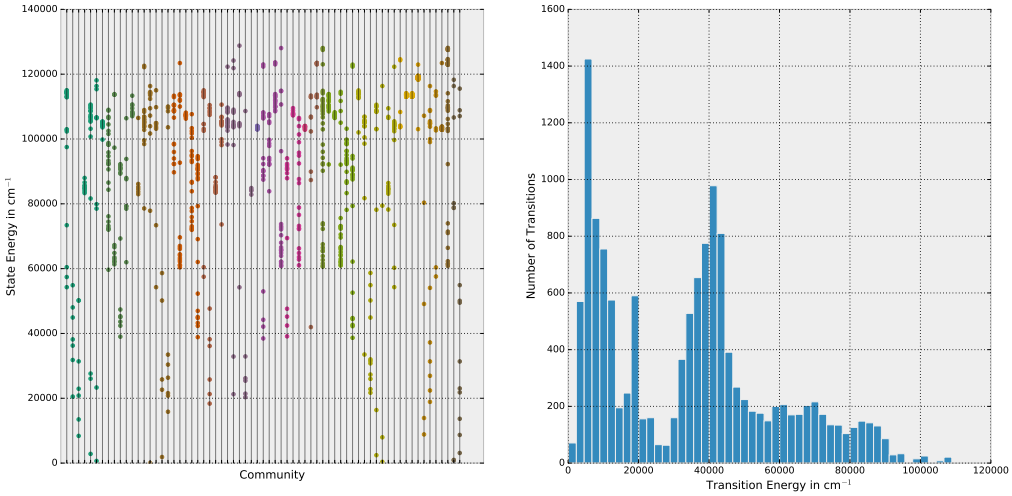


Figure A.7: Energies and Communities in Fe II

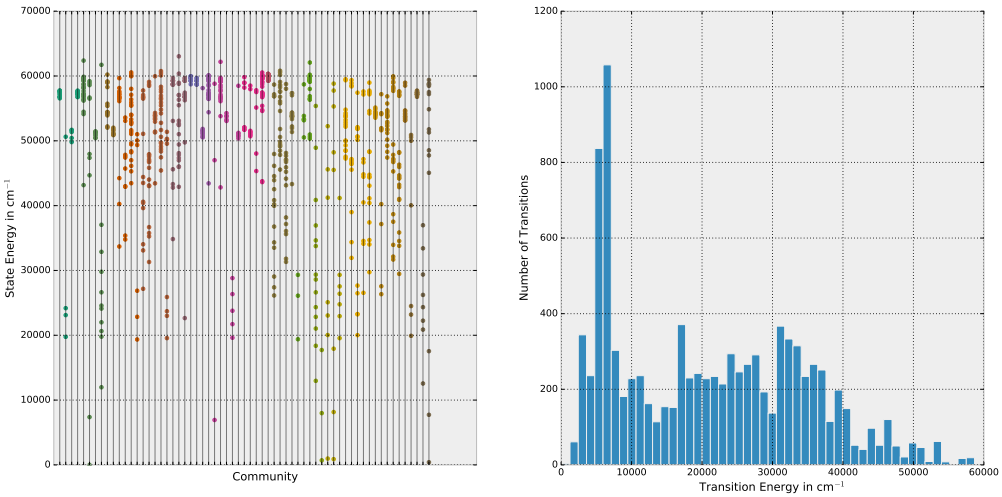


Figure A.8: Energies and Communities in Fe I

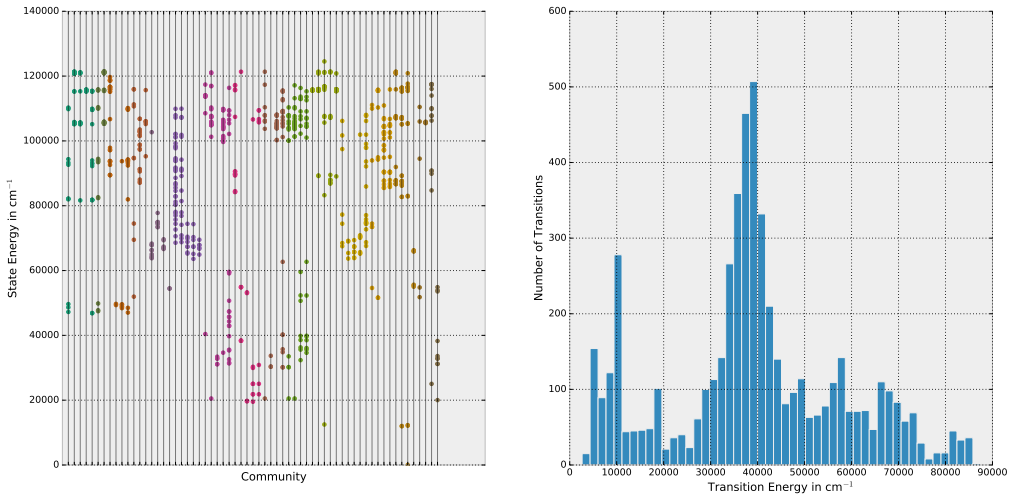


Figure A.9: Energies and Communities in Cr II

B Nested Stochastic Block Model

In this chapter we analyze how the NSBM can be used for link prediction. The general formalism is described in Peixoto [2017], but some modifications are made here due to computational complexity.

According to equation 82 in Peixoto [2017] the probability that the observed network $\mathbf{A}^O$ is modified by the edge set $\delta\mathbf{A}$ is given by:

$$P(\delta\mathbf{A}|\mathbf{A}^O) = \frac{P_{\delta\mathbf{A}}(\delta\mathbf{A}|\mathbf{A}^O + \delta\mathbf{A}) \sum_{\mathbf{b}} P_G(\mathbf{A}^O + \delta\mathbf{A}|\mathbf{b})P(\mathbf{b})}{P(\mathbf{A}^O)} \tag{B.1}$$

Here, $\mathbf{b}$ indicates the different possible community structures of the network, $P_{\delta\mathbf{A}}$ is the distribution modeling the error sources and P_G is the distribution of the possible networks given a certain community structure. This can be brought into a form that is easier to evaluate computationally by some Bayesian statistics and omitting $P(\mathbf{A}^O$ as a constant factor. This is shown in Peixoto [2017].

In addition, we will make some further simplifications: Since we want the likelihood of the link e being part of the network without considering what else is part of the network, we would need to further sum over all edge sets $\delta\mathbf{A} \supset e$. Since this would not be computationally tractable, only $\delta\mathbf{A} = e$ will be considered.

Instead of summing over all possible $\mathbf{b}$ in Peixoto [2017] Monte Carlo sampling is proposed. It seems though as if to properly model the distribution $P(\mathbf{b}|\mathbf{A}^O)$ needs to much sampling in our case. Instead, we will look at the most likely community structures $\mathbf{b}$ and calculate their exact probabilities.

This leads to the following algorithm:

1. Find a likely community structure $\mathbf{b}$ using Peixoto [2018b]
2. Did we already consider $\mathbf{b}$:
   a) Yes: Go back to step 1!
   b) No: Continue!
3. Calculate the entropy S

4. Is the entropy more than highest already found entropy $S_{\max} + 5$
   a) Yes: Go back to step 1!
   b) No: Continue!
5. Calculate the likelihood $p(e)$ for every link e.
6. For each edge add $p(e) \cdot \exp(-(S - S_0))$ to its total likelihood $p_{\text{tot}}(e)$, where S_0 is some offset.

Now, up to a constant factor, $p_{\text{tot}}(e)$ should be a good approximation for the correct likelihood $P(e)$ and will be used as the link prediction score of the NSBM.

C Node Prediction

In this appendix the eigenvalues and eigenvectors of the graph are analyzed with regard to their use for predicting new nodes.

C.1 Influence of Node Removal on the Spectrum

In this section the change of eigenvectors and eigenvalues when nodes are removed at random is analyzed. The relation of a graph to its eigenvalues and eigenvectors has been analyzed in depth in the literature and is known as spectral graph theory. For special graphs (random graphs, regular graphs, scale-free graphs, ...) many results have been derived. Further, many general network properties have been set in relation either to properties of the distribution of eigenvalues or to properties of the largest eigenvectors of adjacency, Laplacian and normalized Laplacian. An example is that a graph is bipartite if and only if for every eigenvalue λ of the adjacency matrix $-\lambda$ is part of its spectrum as well. Literature on spectral graph theory is for example given by Chung and Graham [1997] and Cvetković et al. [1997].

Since these basic properties should not change due to random node removal, the eigenvalues and the important eigenvectors should not change either (or only change in a trivial manor). This assumption needs to be confirmed. We will do that at the example of the thorium II network, but the results generalize to the other spectroscopic networks and some further networks.

C.1.1 Analysis

First, we look at the eigenvalue distribution. The distribution is plotted in figure C.1 before and after the removal of one third of the nodes. The following features can be seen: The shape of the distribution does not change, it just gets squeezed by the node removal. For the largest and smallest eigenvalues this means they get pushed to the center. This is reasonable considering the largest eigenvalue is closely linked to the maximum degree. Node removal leads to a smaller maximum degree,

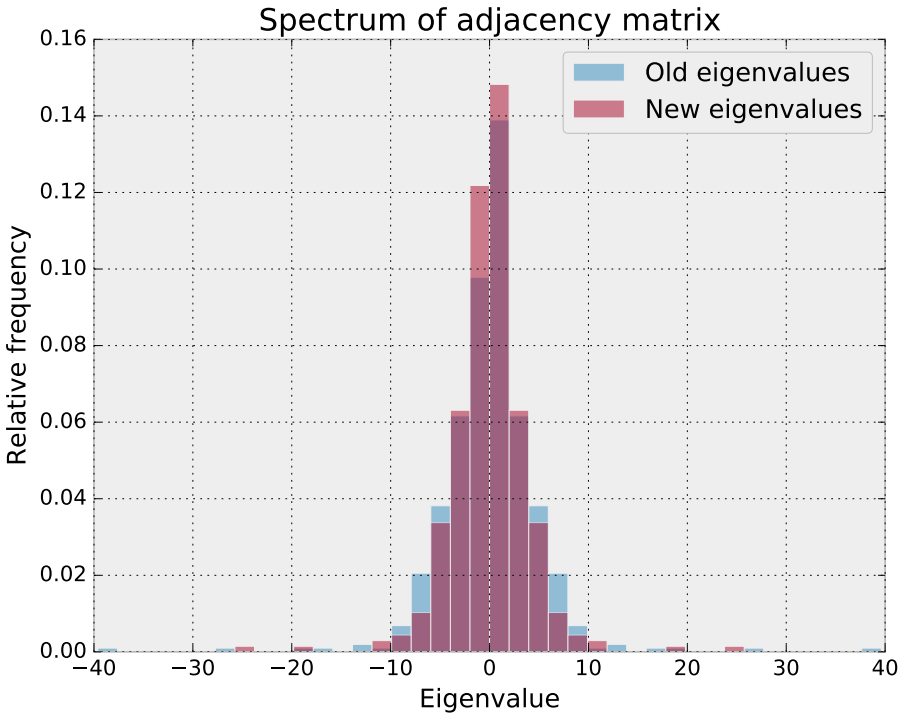


Figure C.1: Histogram of the eigenvalues of the thorium II graph before and after removing one third of the nodes. The general shape of the histogram does not changed, but its variance decreases.

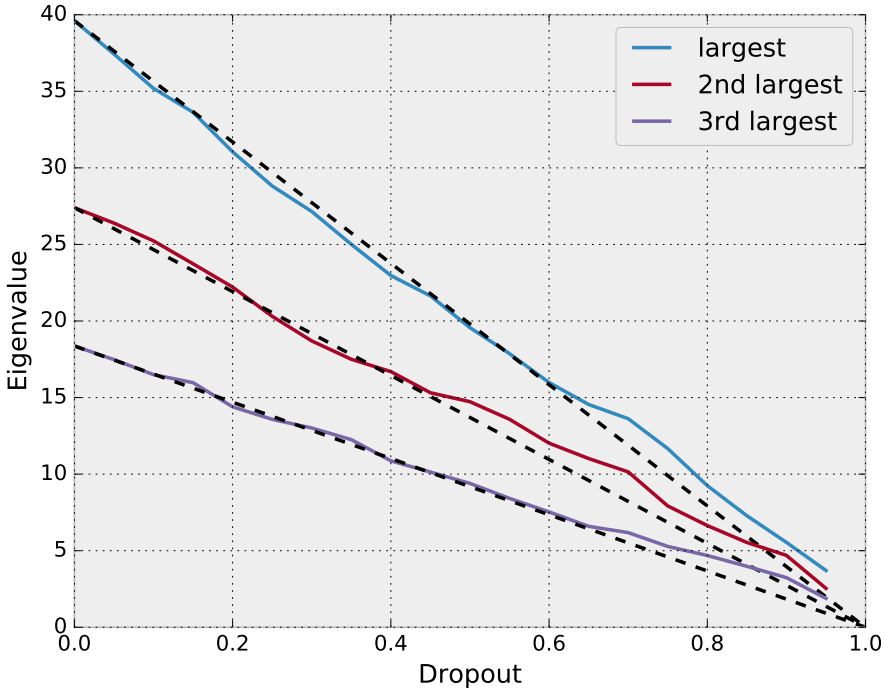
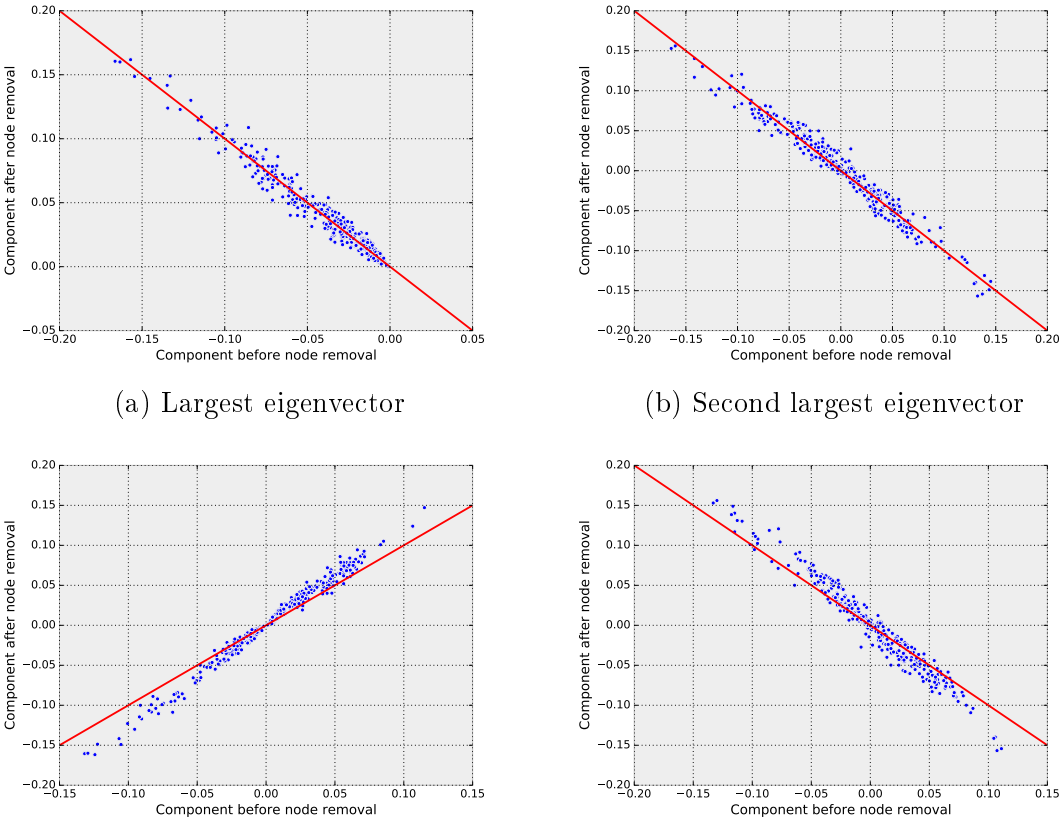


Figure C.2: Scaling of the largest three eigenvalues under random node removal. The dashed lines indicate what scaling by the fraction of nodes would look like.

so the eigenvalue should shrink. It is worth noting here that none of the standard distributions (semi-circle, exponential, polynomial, Gaussian, ...) seems to fit this degree distribution.

In figure C.2 we see that at least up to 40 % node removal the big eigenvalues scale with the number of removed nodes. This holds for the negative eigenvalues, as the negative eigenvalues for bipartite networks are symmetric to the positive ones. For thorium II only dipole transitions were measured, hence the network is bipartite; this cannot change due to node removal. Such scaling corresponds to a constant factor for the entire adjacency matrix and can thus be neglected for relative scores.

Now we analyze how the eigenvectors of the adjacency change upon node removal. For this purpose, we compare all components of the eigenvectors before and after node removal. This is done in figure C.3. To interpret this plot consider: If the nodes eigenvectors were completely unchanged with the exception of removing the components of the deleted nodes, the components would only increase due to normalization. The old components were renormalized such that this effect is already



(a) Largest eigenvector

(b) Second largest eigenvector

(c) Smallest eigenvector

(d) Second smallest eigenvector

Figure C.3: Components of some eigenvectors of the adjacency before and after removing one third of the nodes. Only the components of the kept nodes are shown. Both eigenvectors are normalized such that the plotted components have length one. They all lie roughly on the diagonal, so the eigenvectors are mostly unchanged.

taken care of. Thus all components would just lie on the diagonal (red line). This diagonal can have slope ± 1 as the absolute sign of the eigenvector is undetermined and thus a sign flip between before and after has no meaning. In figure C.3 we can see that this is indeed the case for three of the eigenvectors.

There are only minor changes in the individual components. For the small eigenvectors the slope increases slightly compared to the diagonal. A further analysis of this feature would be beyond the scope of this thesis. In addition the components seem to fluctuate statistically around the diagonal. There are many possible reasons for such fluctuations, the simplest one being that the adjacency matrix still needs to consist of zeros and ones, which is not ensured if all eigenvector components stay the same. Further all nodes are individual, so removing one node should have a slight influence beyond cutting the component. For each individual component the impact of the slope change and statistical fluctuations seem about equally large, so neglecting them will have a minor impact.

For eigenvectors with eigenvalues close to zero the assignment of before and after becomes less clear, as the spaces between the eigenvalues decrease. That means what might be the fifth largest eigenvalue originally might become the sixth largest eigenvalue and vice versa. Furthermore, as argued in 3.2 the effect of the small eigenvalues and their eigenvectors on the adjacency matrix is minor, so not considering them here should not be a big deal.

C.1.2 Conclusion

In this section we found that random node removal leads to a simple change in the spectrum: The individual eigenvalues scale with the size of the network, but the general shape of the eigenvalue-spectrum does not change. The important eigenvectors corresponding to these eigenvalues remain mostly unchanged up to node removal. For both eigenvalues and eigenvectors additional statistical fluctuations have been observed.

C.2 How the Eigenvector Method Modifies the Spectrum

In this section we analyze how the node prediction method we proposed based on eigenvectors influences the spectrum. First some naming conventions: The fraction

of nodes to be added is called p. The adjacency matrix before adding new nodes will be called $\mathbf{A}$ with eigenvalues λ_i and eigenvectors $\vec{v}_i$. The new matrix generated is called $\tilde{\mathbf{A}}$ and the modified eigenvectors used in the process will be called $\tilde{\vec{v}}_i$. These are not the eigenvectors of $\tilde{\mathbf{A}}$, as the vectors $\tilde{\vec{v}}_i$ are not orthogonal to each other. With the definition $\varepsilon_{ij} = \tilde{\vec{v}}_i \cdot \tilde{\vec{v}}_j$ for $i \neq j$ we find:

$$\tilde{\mathbf{A}}\tilde{\vec{v}}_i = \left(\sum_j \lambda_j \tilde{\vec{v}}_j \tilde{\vec{v}}_j^T\right) \tilde{\vec{v}}_i \tag{C.1}$$

$$\approx (1+p)\lambda_i \tilde{\vec{v}}_i + \sum_{j \neq i} \lambda_j \varepsilon_{ij} \tilde{\vec{v}}_j \tag{C.2}$$

Here, we used $\tilde{\vec{v}}_i \cdot \tilde{\vec{v}}_i \approx 1+p$, as each squared component should be on average equally large.

Further ε_{ij} should be small, as positive and negative components in the scalar product should roughly average out. This means $\tilde{\vec{v}}_i$ are almost eigenvectors of $\tilde{\mathbf{A}}$. The other way around This means the true eigenvectors of $\tilde{\mathbf{A}}$ should be almost $\tilde{\vec{v}}_i$, which is one of the restrictions we had.

Showing that the eigenvalues scale by a constant factor plus small deviations is not as easy and will not be done at this point. The core argument should be that $\sum_i \varepsilon^2$ and $\sum_{i,j} \varepsilon^3$ are small for appropriate indices for ε. ε should scale like $\sqrt{\frac{p}{N}}$, and if it can be treated as a random variable each sum can be approximated by a product with $\sqrt{N}$. Hence the eigenvalues should only deviate little ($\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$) from a constant factor. Here, it shall suffice to say that the method proposed works well for predicting new nodes.

D Lists

D.1 List of Figures

E Bibliography

E. Abbe. Community detection and stochastic block models: recent developments. *arXiv preprint arXiv:1703.10146*, 2017.

A.-L. Barabási et al. *Network science*. Cambridge university press, 2016.

B. H. Bransden, C. J. Joachain, and T. J. Plivier. *Physics of Atoms and Molecules*. Pearson Education. Prentice Hall, 2003. ISBN 9780582356924.

F. R. K. Chung and F. C. Graham. *Spectral graph theory*. Number 92. American Mathematical Soc., 1997.

A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

A. Clauset, C. Aicher, A. Z. Jacobs, and A. Clauset. The weighted stochastic block model. *arXiv*, (November):1–6, 2013.

R. D. Cowan. *The theory of atomic structure and spectra*. Number 3. Univ of California Press, 1981.

A. G. Császár, T. Furtenbacher, and P. Árendás. Small molecules—big data. *The Journal of Physical Chemistry A*, 120(45):8949–8969, 2016. doi: 10.1021/acs.jpca.6b02293.

D. M. Cvetković, P. Rowlinson, and S. Simic. *Eigenspaces of graphs*, volume 66. Cambridge University Press, 1997.

R. Eyal, A. Rosenfeld, S. Sina, and S. Kraus. Predicting and identifying missing node information in social networks. *ACM Trans. Knowl. Discov. Data*, 8(3): 1–35, 2013. doi: 10.1145/2536775.

S. Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

S. Fortunato and D. Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, 2016.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

R. Guimerà and M. Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078, 2009.

A. Halu, L. Ferretti, A. Vezzani, and G. Bianconi. Phase diagram of the bose-hubbard model on complex networks. *EPL (Europhysics Letters)*, 99(1):18001, 2012.

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

O. A. Herrera-Sancho, N. Nemitz, M. V. Okhapkin, and E. Peik. Energy levels of th+ between 7.3 and 8.3 ev. *Physical Review A*, 88(1):012512, 2013.

D. Hric, T. P. Peixoto, and S. Fortunato. Network structure, metadata, and the prediction of missing nodes and annotations. *Phys. Rev. X*, 2016. doi: 10.1103/PhysRevX.6.031038.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1): 193–218, 1985. doi: 10.1007/BF01908075.

O. Jitrik and C. F. Bunge. Transition Probabilities for Hydrogen-Like Atoms. *Journal of Physical and Chemical Reference Data*, 33:1059–1070, December 2004. doi: 10.1063/1.1796671.

M. Kim and J. Leskovec. The Network Completion Problem: Inferring Missing Nodes and Edges in Networks. *SIAM Int. Conf. Data Min.*, pages 47–58, 2011. doi: 10.1137/1.9781611972818.5.

A. Kramida, Yu. Ralchenko, and J. Reader and NIST ASD Team. NIST Atomic Spectra Database (ver. 5.5.6), [Online]. Available: https://physics.nist.gov/asd [2017, April 9]. National Institute of Standards and Technology, Gaithersburg, MD., 2018.

N. Kulvelis, M. Dolgushev, and O. Mülken. Universality at breakdown of quantum transport on complex networks. *Physical review letters*, 115(12):120602, 2015.

S. Lackner, A. Spitz, M. Weidemüller, and M. Gertz. Efficient anti-community detection in complex networks. In *Proceedings of the 30th International Conference on Scientific and Statistical Database Management*, SSDBM '18, pages 16:1–16:12, New York, NY, USA, 2018. ACM. doi: 10.1145/3221269.3221289.

Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. Facing the cold start problem in recommender systems. *Expert Syst. Appl.*, 41(4 PART 2):2065–2073, 2014. doi: 10.1016/j.eswa.2013.09.005.

C. Lovis and F. Pepe. A new list of thorium and argon spectral lines in the visible. *Astronomy & Astrophysics*, 468(3):1115–1121, 2007.

L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Phys. A Stat. Mech. its Appl.*, 390(6):1150–1170, 2011. doi: 10.1016/j.physa.2010.11.027.

L. Lü, L. Pan, T. Zhou, Y.-C. Zhang, and H. E. Stanley. Toward link predictability of complex networks. *Proceedings of the National Academy of Sciences*, 112(8): 2325–2330, 2015. doi: 10.1073/pnas.1424644112.

M. E. J. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

M. E. J. Newman. *Networks: An Introduction*. 2010. doi: 10.1093/acprof:oso/9780199206650.001.0001.

E. Peik and C. Tamm. Nuclear laser spectroscopy of the 3.5 ev transition in th-229. *EPL (Europhysics Letters)*, 61(2):181, 2003.

T. P. Peixoto. Bayesian stochastic blockmodeling. *ArXiv e-prints*, May 2017.

T. P. Peixoto. Nonparametric weighted stochastic block models. *Phys. Rev. E*, 97 (1), 2018a. doi: 10.1103/PhysRevE.97.012306.

T. P. Peixoto. graph-tool, 2018b. URL https://graph-tool.skewed.de/.

S. L. Redman, G. Nave, and C. J. Sansonetti. The spectrum of thorium from 250 nm to 5500 nm: Ritz wavelengths and optimized energy levels. *The Astrophysical Journal Supplement Series*, 211(1):4, 2014.

M. S. Safronova, U. I. Safronova, and C. W. Clark. Relativistic all-order calculations of th, th+, and th 2+ atomic properties. *Physical Review A*, 90(3):032512, 2014.

L. H. Son. Dealing with the new user cold-start problem in recommender systems: A comparative review. *Inf. Syst.*, 58:87–104, jun 2016. doi: 10.1016/j.is.2014.10.001.

M. A. Valdez, D. Jaschke, D. L. Vargas, and L. D. Carr. Quantifying complexity in quantum phase transitions via mutual information complex networks. *Physical review letters*, 119(22):225301, 2017.

J. Wei, J. He, K. Chen, Y. Zhou, and Z. Tang. Collaborative filtering and deep learning based recommendation system for cold start items. *Expert Syst. Appl.*, 69:1339–1351, 2017. doi: 10.1016/j.eswa.2016.09.040.

W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.

R. Zalubas and C. H. Corliss. Energy levels and classified lines in the second spectrum of th(th ii). *J. Res. NBS A.*, 78(2):163–246, 1974.

Acknowledgements

At this point I want to thank everyone who helped in the process of this thesis and without whom this thesis would not have been possible. A first thank you goes to Matthias Weidemüller, who supervised this thesis and provided many helpful insights into atomic physics and of the organization of the thesis.

A big thanks goes further to Julian Heiss and Armin Kekić. They were part of much research that has been done in this thesis. Many helpful discussions with them stimulated the working process and many crucial results were discovered in collaboration with them.

Further I want to thank Michael Gertz, Sebastian Lackner and especially Andreas Spitz for many helpful insights into network theory. They provided the necessary expertise about network methods and the state of research in network science.

I further want to acknowledge the members of the group of Matthias Weidemüller, with whom many fruitful discussions were done. Special thanks goes here to Benjamin Claßen and Klaus Kades who were always open to listen to new ideas and give their opinion.

Last but not least, I also want to thank Julian Heiss, Klaus Kades and Nandita Raman, who proofread this thesis.

Erklärung:

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

Heidelberg, den (Datum) .....