SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
SPECTROSCOPY
QUANTUM CHEMISTRY

Received
5 December 2013

Accepted
27 March 2014

Published
11 April 2014

Correspondence and requests for materials should be addressed to A.G.C. (csaszar@chem.elte.hu)

Simple molecules as complex systems

Tibor Furtenbacher[1,2], Péter Árendás[2,3], Georg Mellau[4] & Attila G. Császár[1,2]

[1]Laboratory of Molecular Structure and Dynamics, Institute of Chemistry, Eötvös Loránd University, H-1117 Budapest, Pázmány Péter sétány 1/A, Hungary, [2]MTA-ELTE Research Group on Complex Chemical Systems, H-1518 Budapest 112, P.O. Box 32, Hungary, [3]Department of Algebra and Number Theory, Institute of Mathematics, Eötvös Loránd University, H-1518 Budapest 112, P.O. Box 120, Hungary, [4]Physikalisch-Chemisches Institut, Justus-Liebig-Universität Giessen, Heinrich-Buff-Ring 58, D-35392 Giessen, Germany.

For individual molecules quantum mechanics (QM) offers a simple, natural and elegant way to build large-scale complex networks: quantized energy levels are the nodes, allowed transitions among the levels are the links, and transition intensities supply the weights. QM networks are intrinsic properties of molecules and they are characterized experimentally via spectroscopy; thus, realizations of QM networks are called spectroscopic networks (SN). As demonstrated for the rovibrational states of $H_2{}^{16}O$, the molecule governing the greenhouse effect on earth through hundreds of millions of its spectroscopic transitions (links), both the measured and first-principles computed one-photon absorption SNs containing experimentally accessible transitions appear to have heavy-tailed degree distributions. The proposed novel view of high-resolution spectroscopy and the observed degree distributions have important implications: appearance of a core of highly interconnected hubs among the nodes, a generally disassortative connection preference, considerable robustness and error tolerance, and an "ultra-small-world" property. The network-theoretical view of spectroscopy offers a data reduction facility via a minimum-weight spanning tree approach, which can assist high-resolution spectroscopists to improve the efficiency of the assignment of their measured spectra.

High-resolution molecular spectroscopy is one of the high-end analytical tools which can be used to obtain detailed chemical information about complex natural systems. These systems include the earth's atmosphere, where spectroscopy helps to understand the greenhouse effect, and astronomical bodies of our universe, where spectroscopy helps, among other things, to answer principal questions concerning life on earth. The extensive spectroscopic data required by related modelling efforts have been consolidated into information systems[1–11]. The data deposited in these information systems traditionally come from a large number of high-resolution experimental investigations. Experiments are usually done by different groups employing different techniques in different regions of the spectrum, resulting in a broad range of data accuracy. The relative accuracy of transition frequencies detected in the lab ranges from 10^{-5} to 10^{-10}, while for transition intensities it is only 10^{-2}. As to theory, in the fourth age of quantum chemistry[12] it is possible to determine accurate high-resolution spectroscopic data and spectra[13,14]. To satisfy the demand of modellers, for a number of small molecules nearly complete first-principles linelists have been computed[15]. These lists contain from thousands to millions of entries in the form of rotational-vibrational-electronic energies and transitions and their most important characteristics (*e.g.*, quantum numbers, symmetries, and intensities).

Although high-resolution spectroscopic experiments yield highly accurate data, at the same time these data are highly incomplete. For example, the 5 000 experimental eigenenergies reported by Mellau[16–18] are complete up to 7 000 cm^{-1} above the HCN ground state, yet they cover only 98 vibrational states. The 25 000 rovibrational states determined in these high-resolution infrared emission studies correspond only to 15% of the vibrational states up to isomerization. When compared with experimental data, *ab initio* linelists show the following important characteristics: while the relative accuracy of the *ab initio* energy levels is 10 to 10 000 times worse than that of typical experimental data, most of the transition intensities have accuracies similar to experimental data. The striking disparity between the accuracy and the number of first-principles computed and experimentally measured energy levels and transitions and the fact that in many cases *ab initio* intensities may directly be used for high resolution analyses leads to the conclusion that for the foreseeable future one should consider the combination of experimental and *ab initio* information to satisfy the needs of modellers, who often require nearly complete high-resolution (line by line) spectroscopic data[19]. In turn, this conclusion leads immediately to questions how results of the various experiments should be viewed, how experimental and theoretical data could be unified, how *ab initio* data may be used to simplify the assignment of measured spectra, and how to build the most dependable information systems containing line-by-line spectroscopic data.

We believe that to obtain the best answers to these questions one should consider the energy levels and the spectroscopic transitions of a molecule from the point of view of graph theory. Thus, earlier we introduced the

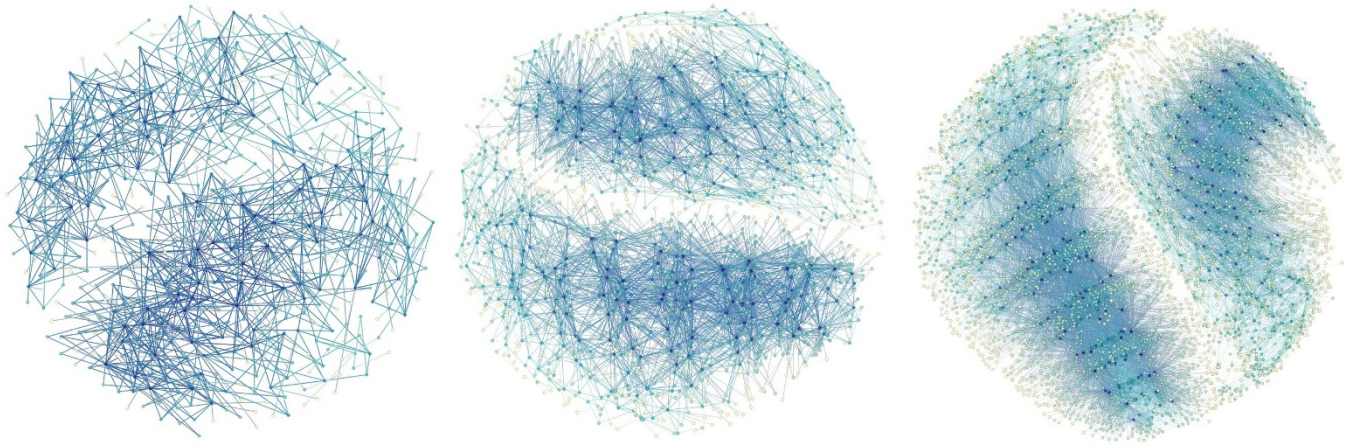

Figure 1 | Visual representation of the first-principles spectroscopic networks of $H_2{}^{16}O$ in absorption with an intensity cut-off of 10^{-20}, 10^{-22}, and 10^{-24} cm molecule^{-1}, from left to right, with clearly visible ortho and para components and buildup of hubs.

concept of spectroscopic networks (SN)[20–24], where quantized energy levels are the nodes (vertices) and allowed transitions among the levels are the links (edges) of a graph (see Fig. 1). SNs are considered to be an intrinsic property of molecular systems, though characteristics of SNs can be slightly different based on how we actually probe these systems experimentally (*e.g.*, in absorption or in emission). SNs provide a convenient representation of the experimental and theoretical data and ways for their most advantageous unification, as well.

In this paper we extend the network-theoretical analysis of SNs and, furthermore, develop novel tools for high-resolution spectroscopy research based on the concept of SNs. We use $H_2{}^{16}O$ as the model system of our present investigation. The SN of the $H_2{}^{16}O$ molecule is chosen for several reasons. Water is the most abundant polyatomic molecule in the Universe. It is present in many different environments and at many different temperatures. Detailed characterization of the spectroscopic properties of this triatomic molecule is needed to understand and predict the greenhouse effect on earth and its spectroscopy is of high astrophysical and astrochemical relevance. Furthermore, $H_2{}^{16}O$ was the subject of a large number of experimental high-resolution spectroscopic studies validated recently[25]. This experimental dataset of $H_2{}^{16}O$, one of the spectroscopically most thoroughly studied molecules, contains 14 319 nodes (energy levels) and 97 868 unique links (transitions)[25]. A high-quality first-principles linelist[26], including energy levels, assignments, transitions, and Einstein A coefficients, is also available for $H_2{}^{16}O$. This computed, so-called BT2 linelist contains altogether 221 097 nodes and 505 806 255 links. Based on the number of nodes and links and the underlying structure one can conclude that even this simple triatomic molecule corresponds to a very complex system if the allowed one-photon transitions among its quantized energy levels are considered.

Spectroscopic networks

A graph G, corresponding to an SN of a molecule, say $H_2{}^{16}O$, is an ordered pair, $G = (L,T)$, where L is the set of energy levels (vertices) and T is a set of transitions (edges), the edges being 2-element subsets of L (see Fig. 1). The number of transitions that emanate from an energy level is called the degree of the level. SNs do not contain loops and since different experiments may measure the same transitions, SNs corresponding to experiments are in fact multigraphs. First-principles SNs are, on the other hand, simple graphs. SNs contain a large number of cycles of widely differing size. In SNs non-negative transition intensities, different for different experimental techniques, are assigned to edges as weights. In summary, SNs are large, finite, weighted, and rooted graphs.

Construction of a first-principles SN goes through the following steps: (1) take all (available) energy levels for the given molecule as nodes; (2) use the quantum chemical selection rules appropriate for the molecule and the experiment to link the nodes; and (3) add the intensities as weights to the links based on the type of experiment and the chosen temperature. The number of links in the graph built is naturally much smaller than all the possible links between the nodes. Consequently, the corresponding adjacency matrix is extremely sparse. In the particular case of $H_2{}^{16}O$, consideration of nuclear spins results in two distinct connection schemes. In the language of graph theory these are components of a network. The two principal components (PC) correspond to the two nuclear spin isomers (usually called "ortho" and "para") of $H_2{}^{16}O$ and both have unique roots. Selection rules cause the two PCs of the SN of $H_2{}^{16}O$ to be bipartite graphs. This interesting fact explains why only even-numbered cycles exist in the SN of $H_2{}^{16}O$ and of molecules of a similar nature[27].

Measurements map only a very limited part of an SN and yield a graph called A_m. The intensity of the transitions is responsible for the incompleteness of A_m as below a certain intensity it is impossible to detect a transition in a given type of experiment. Using the intensity as a cut-off parameter, a series of model networks can be constructed from the complete SN built upon the BT2 linelist[26]. We used the following cut-off parameters to construct model networks for the examination of the evolution of one-photon absorption SNs: 10^{-20}, 10^{-22}, 10^{-24}, 10^{-26}, and 10^{-28} cm molecule^{-1} (see Fig. 1 for a visual

Table 1 | General properties of the spectroscopic networks considered for $H_2{}^{16}O$

Quantity	A_{20}	A_{22}	A_{24}	A_{26}	A_{28}	A_m
intensity cut-off	10^{-20}	10^{-22}	10^{-24}	10^{-26}	10^{-28}	measured
number of nodes	547	1 952	5 815	15 603	44 843	18 572
number of links	1 238	7 288	31 283	115 886	397 147	98 927
$S(G)$	0.617	0.429	0.284	0.182	0.091	0.130
$r(G)$	−0.199	−0.356	−0.420	−0.443	−0.469	−0.202
diameter, d	19	23	28	31	34	44
average path length	5.9	6.2	6.5	6.8	7.1	10.7

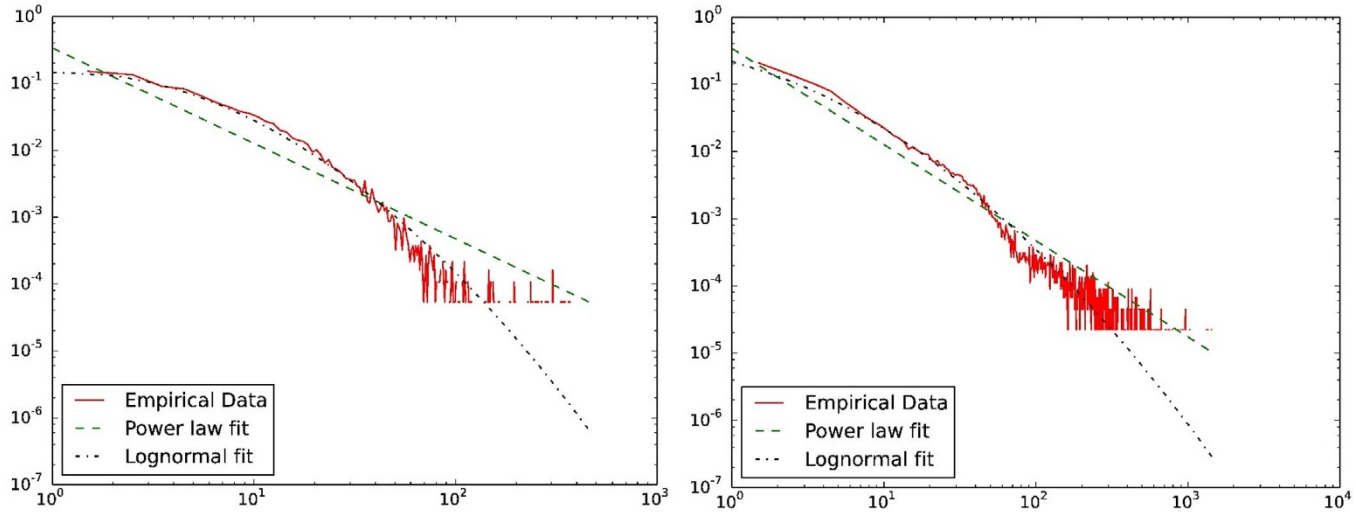

Figure 2 | Distribution of links among nodes given as log-log size–frequency [log$k$ − log$P(k)$] plots for the measured (A_m, left panel) and a first-principles (A_{28}, right panel) spectroscopic network of one-photon absorption transitions for $H_2{}^{16}O$.

representation of three of the first-principles model SNs and Table 1 for details about these SNs, including the number of nodes and links they possess). To emphasize that these SNs belong to absorption, the corresponding graphs are called $A_{20} - A_{28}$.

Floating components (FC), those which do not connect to the roots of PCs, arise frequently in measurements. Since no known transitions exist between the two PCs of the rovibrational SN of $H_2{}^{16}O$, the absolute energy of the higher-energy root, set to a relative energy of zero by definition, can be determined only from an outside source, hindering the high-accuracy absolute determination of all measured energy levels. Artificial transition energies connecting roots of SNs may be called "magic numbers". The traditional route to obtain them is provided by highly accurate model Hamiltonians. A network-theoretical possibility is to take advantage of omnipresent degeneracies of certain higher-energy rovibrational levels in the two PCs, which can be identified straightforwardly by fourth-age[12] variational nuclear-motion computations. These degeneracies are able to connect the distinct components via zero-energy artificial transitions. This was done in Ref. 25 for $H_2{}^{16}O$ and in Ref. 28 for $D_2{}^{16}O$ with the comforting result that the network-theoretical and model Hamiltonian approaches yield the same magic number.

Degree distributions

For many observables there is a typical mean value they cluster around. As to SNs, where the number of experimentally measured links is about an order of magnitude larger than the number of nodes[25,27,29–32], the question is whether there is a mean value for the number of transitions that an "average" energy level has. To answer this question one needs to investigate the distribution of the links among the nodes.

Fig. 2 depicts the size–frequency [log$k$ − log$P(k)$] plots for the A_m and A_{28} SNs of $H_2{}^{16}O$. One can find a very broad distribution and, apart from the very low and very high k part, a reasonably linear relationship in both cases. As detailed in the Methods section, an elaborate search has been performed to estimate the form of the underlying discrete degree-distribution functions of these and the other model SNs. The search included a power-law form of $P(k) \propto k^{-\gamma}$, where γ is the scaling index, as well as exponential and log-normal forms. The analyses indicate a definitely heavy-tailed and, after constraining k to the middle range, a power-law-like behavior with a scaling index of about 2 (Table 2, vide infra). As found for many complex networks[33–35], it is not possible to distinguish between the power-law and the log-normal distributions but the exponential distribution is definitely not compatible with the data. The observed heavy-tailed distribution is one of the most important overall characteristics of SNs and it seems to be generally valid for the PCs of SNs[23].

Whether the degree distribution follows a power law or it is just simply top heavy, the degree distribution functions obtained suggest that SNs are characterized by hubs, *i.e.*, a small number of nodes with a large number of connections. As expected, the most important hubs in a room-temperature absorption spectrum are on the ground vibrational state, (0 0 0), where (v_1 v_2 v_3) are approximate vibrational quantum numbers corresponding to symmetric stretch, bend, and antisymmetric stretch, respectively. For A_m the hubs are as follows: $J_{KaKc} = 6_{34}$, 5_{23}, and 4_{23}, with 458, 455, and 447 links, respectively[25], where J_{KaKc} is the standard rigid-rotor-type quantum number notation applied for asymmetric top molecules, such as $H_2{}^{16}O$. In the A_{28} SN the energy levels with the largest number of transitions are 6_{34}(1487), 5_{23}(1433), and 6_{25}(1431), where the number of links is given in parentheses. Remarkably, the two largest hubs coincide, proving how extensive the experimental investigations are for $H_2{}^{16}O$. Note that the most important hub for $HD^{16}O$ in absorption is also the (0 0 0)6_{34} level[23].

To investigate the hubs of SNs further we determined an SN corresponding to emission created from the first-principles BT2 linelist with an intensity cut-off of 10^{-20} cm molecule^{-1} at 1650 K, which could be called E_{20}. In emission the hubs with the largest number of connections belong to different vibrational states, they are the (0 2 0)9_{63}, (0 0 1)6_{33}, and (0 1 0)10_{38} levels with 102, 101, and 100 links, respectively. The most important hubs in absorption appear to be important hubs in emission but the reverse is obviously not true.

Detailed comparison of the connectivity of measured and first-principles hubs helps to determine the "weakest", least well determined hubs within A_m. This allows the design of new experiments

Table 2 | Parameters for the best power-law models fitted to the SNs of $H_2{}^{16}O$

network	scaling index	k_{min}	p(KS)
A_{22}	2.11	4	0.1060
A_{24}	2.13	8	0.1867
A_{25}	2.15	10	0.2853
A_{26}	2.16	14	0.0460
A_{28}	2.10	17	2.56e-09
A_{30}	2.47	6	2.46e-17
A_{40}	2.83	54	~0

which help to determine a more accurate and robust experimental description of the SN with a minimum amount of effort.

One can also ask the question whether the hubs with the largest number of links take part in the most intense transitions. The answer is a clear no. The 6_{34}, 5_{23}, and 4_{23} pure rotational energy levels take part in the 16th, 18th, and 13th most intense rovibrational absorption transitions, respectively. *Vice versa*, the two energy levels taking part in the most intense transition are only 69th and 89th in the list of hubs based on the number of connections.

Complexity measures

Complexity of a graph G can be assessed by several metrics[35–39]. Three of them, $C(G)$, $S(G)$, and $r(G)$ have been investigated in this study (see Table 1).

The local clustering coefficient, $C(G)$[38], quantifies how close local graphs are to being a complete graph. This metric cannot be used for the bipartite PCs of the model SNs of $H_2{}^{16}O$ as bipartite graphs do not contain odd-numbered cycles such as triangles.

A second metric is the structural metric (s-metric) with the corresponding $S(G)$ value[39] (see the Methods section for details). The $S(G)$ values of the different networks investigated are collected in Table 1.

As shown by Newman[36], social networks seem to show "assortative mixing", *i.e.*, their high-degree vertices preferentially attach to other high-degree vertices. On the contrary, technological and biological networks tend to show[36] "disassortative mixing", *i.e.*, their high-degree vertices attach to low-degree ones. A graph assortativity measure is the Pearson correlation coefficient, $r(G)$[39]. The $r(G)$ values for the first-principles and measured SNs investigated are given in Table 1. For details see the Methods section.

Ordinarily[36,37], one expects a large value of $S(G)$ to be associated with a large positive $r(G)$ value. As seen in Table 1, the $S(G)$ and $r(G)$ values decrease when the intensity cut-off parameter of the first-principles SNs is decreased. This unusual behavior can be rationalized once the evolution of the underlying SNs is understood. If we examine the smallest model SN, A_{20} (see the leftmost panel of Fig. 1 for its visual representation), we find that it contains only two components (it would not be surprising if the energy levels involved in the largest intensity lines would produce several components but this is not the case here). In these two components, containing the most intense transitions, the likelihood of connections among high-degree nodes (hubs) is high; in other words, their eigenvalue centrality[37] is high. This is the reason why the $S(G)$ value is relatively large, while $r(G)$ is close to zero. While the $r(G)$ value of A_{20} is negative, the corresponding large $S(G)$ value indicates that this graph is disassortative with hubs showing an assortative behavior. This means that in A_{20} hubs do like to connect to each other but each hub has many connections to low-degree nodes. Investigating the other SNs we can make another interesting and important observation: the nodes characterized as hubs do not change with the cut-off parameter. Of the first 100 hubs of the model A_{20} and A_{28} SNs 98 are common, meaning that the hubs already appear in the smallest SN and hubs remain hubs when the SN is enlarged. When increasing the size of the SN by decreasing the intensity cut-off parameter, the number of low-degree nodes increases substantially and the ratio of the connections among high-degree nodes to that of high-low connections decreases. This is the reason why the $S(G)$ values show a decreasing tendency when going from A_{20} to A_{28} and the SNs become increasingly disassortative. Note also how nicely the experimental SN, A_m, fits this picture, supporting these findings about SNs.

Small worlds

The small world and ultra-small world properties of graph theory characterize networks where the average path length, defined as the average length of the shortest paths, of two arbitrarily chosen nodes scales as $\sim \log N$ or $\sim \log\log N$, respectively, where N is the number of nodes in the network. Scale-free networks are closer to ultra-small worlds[40]. Heuristically this means that most vertices are within reach via a small number of steps.

The structure resulting from the extreme number of connections within a particular SN can be described efficiently by two numbers, the diameter and the average path length. Of the possible definitions of a diameter we use the one which states that the diameter of a network, $d(G)$, is the maximal shortest path between any two vertices. The diameters and the average path lengths of the SNs studied are given in Table 1. The average path length for the first-principles and measured SNs of $H_2{}^{16}O$ is only about 7, the measured SN has a slightly larger value. The diameter of the first-principles SNs grow as the size of the SN grows but remains at relatively small values. As the data of Table 1 suggest, SNs are ultra-small worlds.

Network vulnerability

A spectroscopic network becomes larger either via new measurements (for an experimental SN) or by a decrease in the intensity cut-off (for a first-principles SN). In either case, the number of transitions increases substantially faster than the number of energy levels, in complete accord with the degree distribution observed. The number of cycles within the network also increases drastically. As a result, SNs appear to be extremely robust.

Robustness of SNs can be ascertained by random removal of nodes[41]. In scale-free networks removal of nodes leads to an increase in the diameter[41]. In SNs, after random removal of 10 to 90% of the nodes, $d(G)$ reflects how the graph fragments and thus provides useful characteristics about SNs. The original diameter of the largest first-principles graph investigated, A_{28}, is 34 (Table 1), and this value does not change until we randomly remove some 95% of the nodes. Then the diameter suddenly drops to 22. The observed robustness of the SN of $H_2{}^{16}O$ can be explained by the nature of the selection rules leading to a bipartite graph and the presence of an assortative core of interconnected hubs. To prove the latter we note that in A_{28} the first 448 hubs, 1% of the nodes, own almost 40% of the links. On one hand, the probability of random removal of hubs is small, on the other hand, if we remove such hubs, another hub "takes over" in the graph, as hubs are 'well connected'. The situation is quite different when we attack the graph, *i.e.*, we remove the high-degree nodes systematically. If we delete the first 200 hubs, 0.45% of the nodes, which have 20.45% of the links, the diameter reduces to 18. The extreme error tolerance is another characteristic property of SNs and this property is somewhat similar to that observed in other complex networks.

Data reduction via SNs

Since high-resolution spectroscopic measurements yield an extreme amount of information, the reduction of the data to manageable size is a basic challenge for the theory of spectroscopy. The standard solution is to use model Hamiltonians with a small number of parameters and least-squares optimize these parameters to represent all the measured data[42]. In a way this means that spectroscopic transitions are converted to parameters yielding energy levels. These parameters allow excellent interpolation but they may fail drastically when used to extrapolate beyond the measured range.

SNs offer another data reduction facility via an inversion of transitions to energy levels. For example, the 500 million transitions of the BT2 linelist can be converted back to about 200 thousand energy levels. This feature of SNs has been exploited in the MARVEL (Measured Active Rotational-Vibrational Energy Levels) procedure[21,22] used, among other applications, to derive the IUPAC spectroscopic database of water isotopologues[25,28,29,31,32].

The best way to reduce the information content of SNs is through the use of weighted spanning trees. By using weighted spanning trees[43], see the Methods section, one can reduce the information contained in the huge number of measured transitions of the

complex A_m network to a relatively small set of energy levels. Each link of A_m has a widely different uncertainty. The network-theoretical view allows to appreciate how cycles, containing a lot of extra information compared to, for example, minimum weight spanning trees, within a component of an SN help to fix the energy levels and tighten their uncertainties.

Assignment of spectra

High resolution spectroscopy is also a science (and art) of quantum number assignment of measured lines and levels. The traditional way of analysing high-resolution experimental spectra is the *a priori* assignment of lines with good and approximate quantum numbers followed by a fitting of the levels via a small number of spectroscopic parameters of a well-designed model Hamiltonian[42]. This type of assignment procedure fails in the case of highly excited rovibrational states and in general when the number of rovibrational transitions exceeds a limit corresponding to an acceptable analysis time. A combined microwave to visible spectrum of any polyatomic molecule is converted to a list of labelled eigenenergies[16–18] in a high-resolution study.

Hereby we advocate a novel protocol for the assignment of spectra based on SNs: detect the lines in a measured high-resolution spectrum leading to the largest number of new energy levels via an investigation of a suitable first-principles SN and assign the transitions with quantum numbers by mapping the *ab initio* line-list onto experimental spectra using graph theory. Taking the negative logarithm of the intensity of the transitions as the weight function for the transitions of the SN, the minimum-weight spanning tree displays the transitions with the largest intensities; thus, it readily identifies the most intense and thus the practically most useful spectral features. An illustration of the concept is provided in Fig. 3.

The proposed method based on graph theory allows the automated and fast conversion of very large experimental datasets into complete eigenenergy lists. These lists are the starting points for the development of theoretical models connecting our physical and chemical view on molecules[18].

Finally, let's create an artificial spectrum, in order to show the utility of the weighted spanning-tree approach. The complete set of 1 916 $H_2{}^{16}O$ rovibrational energy levels up to 7 000 cm^{-1} is known with high-resolution accuracy from a MARVEL study[25]. Based on these energy levels a simulated room temperature absorption spectrum is obtained containing 45 266 allowed transitions with intensities larger than 10^{-28} cm molecule^{-1}. The corresponding minimum-weight spanning tree contains 1 914 transitions, the minimum number of *intense* transitions needed to convert the spectrum back to an energy list. This represents a significant, more than 20-fold reduction in the data. In other words, analysis of only 1 914 intense transitions yields the maximum number of energy levels that can be determined from this spectrum. It is worth adding that out of the 45 266 lines 19 482, an order of magnitude more than minimally needed, have indeed been measured and assigned[25], which is a likely unusually high degree of completeness.

Conclusions

Driven by the need of scientific and engineering applications, complex spectroscopic networks, perhaps as part of active databases[20–24], are expected to become an intrinsic part of the description of the high-resolution spectra of molecules. A good opportunity to advance the field of high-resolution molecular spectroscopy and to turn data into knowledge, as emphasized in the article defining the fourth age of quantum chemistry[12] and confirmed here, is offered via the joint use of accurate experiments, accurate first-principles computations, and efficient mathematical and numerical algorithms provided by, for example, graph and database theory.

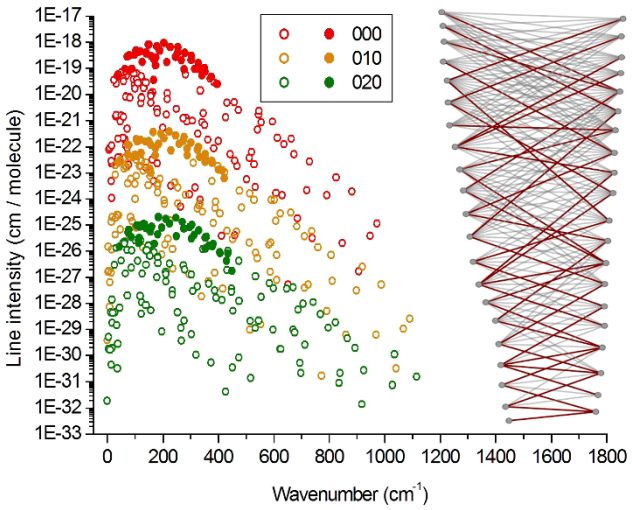

Figure 3 | Rotational spectrum, between 0 and 1100 cm^{-1}, of the first three bands, (0 0 0) (in red), (0 1 0) (in yellow), and (0 2 0) (in green), of para-$H_2{}^{16}O$ for rotational quantum number J less than nine along with the bipartite graph of the transitions, where the spanning tree of the transitions is indicated by red lines and filled circles.

Methods

An assumption at the beginning of this study was that a power-law distribution would be the best choice for modeling the degree distribution of SNs[23]. The in-depth analysis of the degree distributions of the SNs studied utilized a review article[43] and two codes: igraph [**igraph** is a free software package for creating and manipulating undirected and directed graphs, see http://igraph.sourceforge.net/] and an open-source Python package[44]. The density function of power-law distributions can be written as $P(k) \sim L(k)\, k^{-\gamma}$. This function is undefined for $k = 0$; hence, a suitable k_{min} value must be defined. This k_{min} can be specified by various methods, *e.g.*, choosing a noise threshold value or the minimum value in a given sample. Often the low end of the dataset, which contains small values compared to the whole data, does not follow a power-law behavior. Therefore, one can fit a power-law distribution for each value in the dataset acting as k_{min} and compute the best fit by minimalizing the Kolmogorov–Smirnov (KS) distance, p(KS), between the empirical data and the fitted model. After determining the parameters of the power-law distribution, we analyzed our hypothesis that the best model for the empirical degree distribution is the power-law one by implementing a one-sample KS test. We reject the hypothesis if the p values obtained from the test fall below 0.05. The results are summarized in Table 2.

The KS test results suggest that the optimal fitting model depends heavily on the intensity cut-off value used to create the model SN. We observe that A_{25} is a "sweet spot" graph in the power-law modelling of the first-principles absorption SN of $H_2{}^{16}O$. By using lower absorption intensity cut-offs, one can no longer properly fit a power-law distribution to the dataset.

Note that there are two observations which help to explain the observed behavior. First, as we incorporate transitions with smaller intensities the network does not expand in terms of new vertices but becomes denser. Second, we refer the reader to the section on complexity measures. As seen there, the intensities of transitions involving hubs are generally considerably larger than those of non-hub ones. This observation is responsible for the fact that while the number of edges increases, the new edges do not substantially boost the degree of the hubs.

The normalization constant for discrete power-law distributions is $1/\zeta(\gamma, k_{min})$[44], where $\zeta(s, a)$ stands for the Hurwitz zeta function,

$$\zeta(s,a) = \sum_{k=0}^{\infty} \frac{1}{(k+a)^s} \tag{1}$$

We note that we cannot model the empirical degree distribution of the current measured SN, A_m, with a power-law distribution. The same algorithm as above leads us to a scaling index of 2.66 choosing 16 as the optimal k_{min}. However, the KS test gives a p value of 0.02; thus, we must reject the hypothesis that the dataset was drawn from a power-law distribution.

The s-metric is defined by

$$s = \sum\nolimits_{i,j \in T} d_i d_j, \tag{2}$$

where d_i is the degree of node i. If we introduce s_{max} as

$$s_{\max} = \sum_{i=1}^{N} \frac{d_i^3}{2}, \tag{3}$$

we can define the normalized s-metric used in the text as

$$S(G) = s/s_{\max}. \tag{4}$$

The graph assortativity, $r(G)$, is defined by the Pearson coefficient,

$$r(G) = \frac{\sum_{i,j\in T} \frac{d_i d_j}{l} - \left(\sum_{i,j\in T} \frac{d_i + d_j}{2l}\right)^2}{\sum_{i,j\in T} \frac{d_i^2 + d_j^2}{2l} - \left(\sum_{i,j\in T} \frac{d_i + d_j}{2l}\right)^2}, \tag{5}$$

where l is the number of edges in the graph.

To build a minimum-weight spanning tree from the SNs, we implemented Kruskal's algorithm[45]. For the weight function, the negative logarithm value of the intensities on the edges were used. Admittedly, a more accurate result can be achieved by multiplying the base intensity values by −1 to obtain a weight function. Nevertheless, the differences are within the same order of magnitude and are negligible for practical considerations; therefore, we believe the weight function employed is adequate.

1. Rothman, L. S. The evolution and impact of the HITRAN molecular spectroscopic database. *J. Quant. Spectrosc. Rad. Transfer* **111**, 1565–1567 (2010).
2. Rothman, L. S. *et al.* The HITRAN 2008 molecular spectroscopic database. *J. Quant. Spectrosc. Rad. Transfer* **110**, 533–572 (2009).
3. Rothman, L. S. *et al.* HITEMP, the high-temperature molecular spectroscopic database. *J. Quant. Spectrosc. Rad. Transfer* **111**, 2139–2150 (2010).
4. Jacquinet-Husson, N. *et al.* The 2003 edition of the GEISA/IASI spectroscopic database. *J. Quant. Spectrosc. Rad. Transfer* **95**, 429–467 (2005).
5. Landi, E., Young, P. R., Dere, K. P., Del Zanna, G. & Mason, H. E. CHIANTI – An atomic database for emission lines. XIII. Soft X-ray improvements and other changes: Version 7.1 of the database. *Astrophys. J.* **763**, 86 (2013).
6. Müller, H. S. P., Schlöder, F., Stutzki, J. & Winnewisser, G. The Cologne database for molecular spectroscopy, CDMS: A useful tool for astronomers and spectroscopists. *J. Mol. Struct.* **742**, 215–227 (2005).
7. Müller, H. S. P., Thorwirth, S., Roth, D. A. & Winnewisser, G. The Cologne database for molecular spectroscopy, CDMS. *Astron. Astrophys.* **370**, L49–L52 (2001).
8. Pickett, H. M. *et al.* Submillimeter, millimeter and microwave spectral line catalog. *J. Quant. Spectrosc. Rad. Transfer* **60**, 883–890 (1998).
9. Jacquinet-Husson, N. *et al.* The 2009 edition of the GEISA spectroscopic database. *J. Quant. Spectrosc. Rad. Transfer* **112**, 2395–2445 (2011).
10. Dubernet, M. L. *et al.* Virtual Atomic and Molecular Data Centre. *J. Quant. Spectr. Rad. Transfer* **111**, 2151–2159 (2010).
11. Tashkun, S. A., Perevalov, V. I., Teffo, J.-L., Bykov, A. D. & Lavrentieva, N. N. CDSD-1000, the high-temperature carbon dioxide spectroscopic databank. *J. Quant. Spectrosc. Rad. Transfer* **82**, 165–196 (2003).
12. Császár, A. G. *et al.* The fourth age of quantum chemistry: Molecules in motion. *Phys. Chem. Chem. Phys.* **14**, 1085–1106 (2012).
13. Polyansky, O. L. J. *et al.* High-accuracy ab initio rotation-vibration transitions for water. *Science* **299**, 539–542 (2003).
14. Pavanello, M. *et al.* Precision measurements and computations of transition energies in rotationally cold triatomic hydrogen ions up to the mid-visible spectral range. *Phys. Rev. Lett.* **108**, 023002 (2012).
15. Tennyson, J. & Yurchenko, S. N. ExoMol: molecular line lists for exoplanet and other atmospheres. *Mon. Not. R. Astron. Soc.* **425**, 21–33 (2012).
16. Mellau, G. Ch. Complete experimental rovibrational eigenenergies of HNC up to 3743 cm^{-1} above the ground state. *J. Chem. Phys.* **133**, 164303 (2010).
17. Mellau, G. Ch. Complete experimental rovibrational eigenenergies of HCN up to 6880 cm^{-1} above the ground state. *J. Chem. Phys.* **134**, 234303 (2011).
18. Mellau, G. Ch. Rovibrational eigenenergy structure of the [H,C,N] molecular system. *J. Chem. Phys.* **134**, 194302 (2011).
19. Barber, R. *et al.* ExoMol line lists III: An improved hot rotation-vibration line list for HCN and HNC. *Mon. Not. Royal Astron. Soc.* **437**, 1828–1835 (2014).
20. Császár, A. G., Czakó, G., Furtenbacher, T. & Mátyus, E. An active database approach to complete spectra of small molecules. *Annu. Rep. Comp. Chem.* **3**, 155–176 (2007).
21. Furtenbacher, T., Császár, A. G. & Tennyson, J. MARVEL: measured active rotational-vibrational energy levels. *J. Mol. Spectrosc.* **245**, 115–125 (2007).
22. Furtenbacher, T. & Császár, A. G. MARVEL: measured active rotational-vibrational energy levels. II. Algorithmic improvements. *J. Quant. Spectr. Rad. Transfer* **113**, 929–935 (2012).
23. Császár, A. G. & Furtenbacher, T. Spectroscopic networks. *J. Mol. Spectrosc.* **266**, 99–103 (2011).
24. Furtenbacher, T. & Császár, A. G. The role of intensities in determining characteristics of spectroscopic networks. *J. Mol. Struct.* **1009**, 123–129 (2012).
25. Tennyson, J. *et al.* IUPAC critical evaluation of the rotational-vibrational spectra of water vapor. Part III. Energy levels and transition wavenumbers for $H_2{}^{16}O$. *J. Quant. Spectr. Rad. Transfer* **117**, 29–58 (2013).
26. Barber, R. J., Tennyson, J., Harris, G. J. & Tolchenov, R. N. A high accuracy computed water line list. *Mon. Not. R. Astron. Soc.* **368**, 1087–1094 (2006).
27. Furtenbacher, T., Szidarovszky, T., Fábri, C. & Császár, A. G. MARVEL analysis of the rotational-vibrational states of the molecular ions H_2D^+ and D_2H^+. *Phys. Chem. Chem. Phys.* **15**, 10181–10193 (2013).
28. Tennyson, J. *et al.* IUPAC critical evaluation of the rotational-vibrational spectra of water vapor. Part IV. Energy levels and transition wavenumbers for $D_2{}^{16}O$, $D_2{}^{17}O$, and $D_2{}^{18}O$. *J. Quant. Spectr. Rad. Transfer* DOI: http://dx.doi.org/10.1016/j.jqsrt.2014.03.019 (2014).
29. Tennyson, J. *et al.* A Database of Water Transitions from Experiment and Theory (IUPAC Technical Report). *Pure Appl. Chem.* **86**, 71–83 (2014).
30. Fábri, C. *et al.* Variational quantum mechanical and active database approaches to the rotational-vibrational spectroscopy of ketene. *J. Chem. Phys.* **135**, 094307 (2011).
31. Tennyson, J. *et al.* IUPAC critical evaluation of the rotational-vibrational spectra of water vapor. Part I. Energy levels and transition wavenumbers for $H_2{}^{17}O$ and $H_2{}^{18}O$. *J. Quant. Spectr. Rad. Transfer* **110**, 573–596 (2009).
32. Tennyson, J. *et al.* IUPAC critical evaluation of the rotational-vibrational spectra of water Vapor. Part II. Energy levels and transition wavenumbers for $HD^{16}O$, $HD^{17}O$, and $HD^{18}O$. *J. Quant. Spectr. Rad. Transfer* **111**, 2160–2184 (2010).
33. Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J. & Giles, C. L. Winners don't take all: Characterizing the competition for links on the Web. *Proc. Natl. Acad. Sci.* **99**, 5207–5211 (2002).
34. Newman, M. E. J. Power laws, Pareto distributions, and Zipf's law. *Contemp. Phys.* **46**, 323–351 (2005).
35. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D.-U. Complex networks: Structure and Dynamics. *Phys. Rep.* **424**, 175–308 (2006).
36. Newman, M. E. J. Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208701 (2002).
37. Newman, M. E. J. *Networks* (Oxford University Press, Oxford, 2000).
38. Watts, D. J. & Strogatz, S. H. Collective dynamics of "small-world" networks. *Nature* **393**, 440–442 (1998).
39. Li, L., Alderson, D., Doyle, J. C. & Willinger, W. Towards a theory of scale-free graphs: Definition, properties, and implications. *Intern. Math.* **2**, 431–523 (2005).
40. Cohen, R. & Havlin, S. Scale-free networks are ultrasmall. *Phys. Rev. Lett.* **90**, 058701 (2003).
41. Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
42. Watson, J. K. G. *Vibrational Spectra and Structure* [During, J. R. (ed.), Vol. 6, Chap. 1] (Elsevier, Amsterdam, 1977).
43. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703 (2009).
44. Alstott, J., Bullmore, E. & Plenz, D. powerlaw: a Python package for analysis of heavy-tailed distributions. *PLoS ONE* **9**, e85777 (2014).
45. Kruskal, J. B. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Am. Math. Soc.* **7**, 48–50 (1956).

Acknowledgments

This project was supported by the Hungarian Scientific Research Fund (OTKA NK83583) and by an ERA-Chemistry grant.

Author contributions

A.G.C., T.F. and P.Á. conceived and designed the research described. A.G.C. and G.M. co-wrote the paper with contributions from T.F. and P.Á..

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Furtenbacher, T., Árendás, P., Mellau, G. & Császár, A.G. Simple molecules as complex systems. *Sci. Rep.* **4**, 4654; DOI:10.1038/srep04654 (2014).