Small Molecules—Big Data
Attila G. Császár,∗,‡ Tibor Furtenbacher,† and Péter Árendás‡,*

†Laboratory of Molecular Structure and Dynamics, Institute of Chemistry, Eötvös Loránd University, P.O. Box 32, H-1518 Budapest 112, Hungary
‡MTA-ELTE Complex Chemical Systems Research Group, Pézmány Péter sétány 1/A, H-1117 Budapest, Hungary
§Department of Algebra and Number Theory, Institute of Mathematics, Eötvös Loránd University, P.O. Box 120, H-1518 Budapest 112, Hungary

ABSTRACT: Quantum mechanics builds large-scale graphs (networks): the vertices are the discrete energy levels the quantum system possesses, and the edges are the (quantum-mechanically allowed) transitions. Parts of the complete quantum mechanical networks can be probed experimentally via high-resolution, energy-resolved spectroscopic techniques. The complete rovibronic line list information for a given molecule can only be obtained through sophisticated quantum-chemical computations. Experiments as well as computations yield what we call spectroscopic networks (SN). First-principles SNs of even small, three to five atomic molecules can be huge, qualifying for the big data description. Besides helping to interpret high-resolution spectra, the network-theoretical view offers several ideas for improving the accuracy and robustness of the increasingly important information systems containing line-by-line spectroscopic data. For example, the smallest number of measurements necessary to perform to obtain the complete list of energy levels is given by the minimum-weight spanning tree of the SN and network clustering studies may call attention to “weakest links” of a spectroscopic database. A present-day application of spectroscopic networks is within the MARVEL (Measured Active Rotational—Vibrational Energy Levels) approach, whereby the transitions information on a measured SN is turned into experimental energy levels via a weighted linear least-squares refinement. MARVEL has been used successfully for 15 molecules and allowed to validate most of the transitions measured and come up with energy levels with well-defined and realistic uncertainties. Accurate knowledge of the energy levels with computed transition intensities allows the realistic prediction of spectra under many different circumstances, e.g., for widely different temperatures. Detailed knowledge of the energy level structure of a molecule coming from a MARVEL analysis is important for a considerable number of modeling efforts in chemistry, physics, and engineering.

1. INTRODUCTION
Both graph theory, whose more or less definitive start date is 1735, when Euler found an ingenious solution to the Königsberg bridges problem,1 and (high-resolution) molecular spectroscopy, whose less well-defined origin can be found in works of Fraunhofer, Kirchhoff (incidentally, a native of Königsberg), Brewster, and Bunsen in the first half of the 1800s,2 have a rather long history.3 Around the year 2000 both fields witnessed major shifts in their subjects. Graph theory was extended from the study of small to that of truly large-scale, complex systems.4−12 These complex networks contain millions and easily billions of nodes and links,13−15 their analysis often requires algorithms developed for big data. At about the same time, high-resolution molecular spectroscopy witnessed the emergence of more and more complete lists of molecules,14−32 containing not thousands but millions and eventually billions of entries, preferably including various spectral line parameters. Atomic spectroscopic databases33 are just as important for certain modeling studies and they can be quite large, as well.

Complex networks appear ubiquitously in nature, society, communication, and elsewhere.1 Traditionally, small(er) networks have been examined and interpreted via random graph theory, a mainstay of discrete mathematics developed and popularized by Erdős and Rényi around 1960.34−36 Distribution of the number of edges, emanating from a vertex, of the Erdős–Rényi random graphs have a well-defined, characteristic mean value, often referred to as a scale. However, about two decades later it became widely recognized that most networks of practical interest do not follow the laws characterizing the Erdős–Rényi random graphs.7,34,37,38 The first-neighbor degree distributions of most complex natural networks appear to be free of a scale, as emphasized for physicists by Barabási and Albert.4 Since that pioneering study, a plethora of papers appeared developing the “mathematics” and “physics” of complex networks, both natural and model ones, and finding new occurrences of networks following a degree distribution which can be characterized as scale free.7 Following studies by Barabási and co-workers,5,8−11,13 it became clear that for a faithful representation of complex systems the random network theory of Erdős and Rényi34 has to be superseded by that of scaling, including scale-free (SF), random networks.7

Received: March 3, 2016
Revised: September 16, 2016
Published: September 27, 2016
Complete characterization of high-resolution rotation−vibration−electronic (rovibronic) spectra of a considerable number of molecules, the only quantum systems considered here, starting from the microwave and extending to the ultraviolet, is a prerequisite for modeling and understanding of many processes and phenomena relevant in physics, chemistry, and engineering. Modellers of the atmospheres of planets and cool stars as well as those investigating combustion in rocket exhausts and turbine engines need detailed and accurate temperature-dependent line-by-line information, which not
even the most elaborate spectroscopic measurements can provide (for example, at elevated temperatures). In statistical theories of chemical reaction rates, a central role is assumed by the density of (ro)vibrational states. The experimental determination of the density of states and its integral, the total number of accessible states, is a formidable challenge at high(er) energies for all but the smallest molecules. Recent advances in high-resolution molecular spectroscopy led to a considerable increase in the extent of related experimental spectroscopic information. Some of the data have been deposited, sometimes in a critically evaluated and annotated form, in databases.

As an example, Figure 1 shows the temporal development of our experimental understanding of the rovibrational energy level (and transition) structure of the lowest electronic state of the $\text{H}_2\text{O}$ molecule, based on data collected in ref 42. As usual for graphs, the positioning of the vertices and edges of Figure 1 is completely arbitrary. The overlap of the color-coded rovibrational transitions of Figure 1 show how the number of experimental transitions that have been measured and analyzed grew decade by decade until it reached our present understanding, with about 20000 experimental energy levels deduced from about 200000 experimental transitions. Treatment of the rapidly increasing information and the desire to turn information into knowledge requires sophisticated procedures in the generation, accumulation, validation, handling, visualization, and distribution of spectroscopic data.

The result of a spectroscopic measurement is almost always a set of experimental transition wavenumbers and transition intensities for a given wavenumber range, covering selected parts of the energy eigenstates. It is feasible to publish such an experimental list and compare the results with those of previous reports through effective spectroscopic constants describing these transitions with experimental accuracy. The assignment procedure is sped up, and the results are more clearly embedded in the set of existing data if one can access a database with the complete set of existing transition wavenumbers and energy levels and use the combination differences procedure. However, this has rarely been done in the past. This is a significant problem in high-resolution spectroscopy as there is an unusually large number of inherent interdependencies among the energy levels and the transitions, most of which are ignored in the traditional approach. There are several further notable disadvantages related to the traditional way spectroscopists have been assigning their measured high-resolution spectra. These disadvantages often transform into significant problems when the data are deposited in spectroscopic databases. It is worth reviewing here some of these hindrances of the traditional approach.

There are only a few good quantum numbers characterizing the energy levels involved in the measured transitions. Most of the quantum numbers used to describe the states are approximate ones. Creation of a database where the energy levels are labeled with approximate quantum numbers, independently of how meaningful the quantum numbers are, assures that a unique and complete set of experimental energy levels can be built. Nevertheless, the approximate nature of the quantum numbers may become a significant problem when a large number of scientists contribute to the understanding of the experimental rovibronic states of a molecule, this being the usual case, as they tend to use different, often conflicting approximate quantum numbers (e.g., local vs normal mode labels for the vibrations and the different labeling possibilities for symmetric and spherical tops, especially when contortional motions are allowed). The problem is exacerbated at higher excitations when it usually becomes completely unfeasible to provide physically meaningful labels.

If there are only a few rovibrational transitions measured for a given vibrational parent, construction of a reliable energy level database is a formidable challenge at high energies for all but the smallest molecules. Recent advances in high-resolution molecular spectroscopy led to a considerable increase in the extent of related experimental spectroscopic information. Some of the data have been deposited, sometimes in a critically evaluated and annotated form, in databases.

As an example, Figure 1 shows the temporal development of our experimental understanding of the rovibrational energy level (and transition) structure of the lowest electronic state of the $\text{H}_2\text{O}$ molecule, based on data collected in ref 42. As usual for graphs, the positioning of the vertices and edges of Figure 1 is completely arbitrary. The overlap of the color-coded rovibrational transitions of Figure 1 show how the number of experimental transitions that have been measured and analyzed grew decade by decade until it reached our present understanding, with about 20000 experimental energy levels deduced from about 200000 experimental transitions. Treatment of the rapidly increasing information and the desire to turn information into knowledge requires sophisticated procedures in the generation, accumulation, validation, handling, visualization, and distribution of spectroscopic data.

The result of a spectroscopic measurement is almost always a set of experimental transition wavenumbers and transition intensities for a given wavenumber range, covering selected parts of the energy eigenstates. It is feasible to publish such an experimental list and compare the results with those of previous reports through effective spectroscopic constants describing these transitions with experimental accuracy. The assignment procedure is sped up, and the results are more clearly embedded in the set of existing data if one can access a database with the complete set of existing transition wavenumbers and energy levels and use the combination differences procedure. However, this has rarely been done in the past. This is a significant problem in high-resolution spectroscopy as there is an unusually large number of inherent interdependencies among the energy levels and the transitions, most of which are ignored in the traditional approach. There are several further notable disadvantages related to the traditional way spectroscopists have been assigning their measured high-resolution spectra. These disadvantages often transform into significant problems when the data are deposited in spectroscopic databases. It is worth reviewing here some of these hindrances of the traditional approach.

There are only a few good quantum numbers characterizing the energy levels involved in the measured transitions. Most of the quantum numbers used to describe the states are approximate ones. Creation of a database where the energy levels are labeled with approximate quantum numbers, independently of how meaningful the quantum numbers are, assures that a unique and complete set of experimental energy levels can be built. Nevertheless, the approximate nature of the quantum numbers may become a significant problem when a large number of scientists contribute to the understanding of the experimental rovibronic states of a molecule, this being the usual case, as they tend to use different, often conflicting approximate quantum numbers (e.g., local vs normal mode labels for the vibrations and the different labeling possibilities for symmetric and spherical tops, especially when contortional motions are allowed). The problem is exacerbated at higher excitations when it usually becomes completely unfeasible to provide physically meaningful labels.

If there are only a few rovibrational transitions measured for a given vibrational parent, construction of a reliable effective Hamiltonian is significantly hindered. This may be the case, for example, when the measurement is based on the use of a single narrow laser source. There may certainly be a difference between the accuracy and the precision of the energies when only a small portion of a spectrum is analyzed. Determining the accuracy of measurements processed via the traditional approach may be misleading. Lacking the details provided by a global analysis of all measurements may allow the declaration of much too small uncertainties, the incompatibility of the uncertainties attached to the transitions and the underlying energy levels may become unrealistic.

Finally, in an ideal world with zero uncertainties in the measurement results, knowledge of $N$ connected nonzero rovibrionic energy levels would require $N$ measurements (it is natural to take the lowest energy level as zero). However, as the entries of Table 1 exemplify, the usual spectroscopic practice is far away from this situation: it happens that 10 times as many transitions are measured than energy levels determined. Thus, how to best utilize all the measurement results and how to minimize unnecessary and costly experimental efforts seems to be an overly important task in high-resolution spectroscopy.

These problems are realized immediately when spectroscopic databases are upgraded with the latest information, which happens on a regular basis with, for example, the canonical HITRAN database. Each time a large set of corrections is introduced, they improve the overall quality of the database but it often remains unclear to the user which data are responsible for the previous problems and the corrections. The rapid growth of measured transitions and the much less rapid evolution of our knowledge about the energy level set can also be traced back to shortcomings of the traditional, serial approach to high-resolution spectra and spectroscopy.
Considering all the interdependencies of all the measurements all at the same time seems to be the answer to almost all of the problems mentioned. This provides a way to an improved error and uncertainty analysis, a much improved facility to determine the overall accuracy of the complete set of measurements, a clear recipe to provide accurate uncertainties, covariances, and provenances, while allowing us to utilize all the measurements that have dependable uncertainties independent of whether they are small or large.

It is the belief of the authors that in search of useful novel tools for validating and improving energy levels and transitions within spectroscopic line lists, as well as improving the understanding of the underlying experiments, network theory and its sophisticated polynomial algorithms offer interesting and highly useful possibilities, some of which are described and explored below.

This belief is based on the fact that for individual molecules quantum mechanics (QM) offers a simple, natural, and elegant way to build large-scale networks. The QM networks are made up of energy levels, forming nodes (vertices), while the allowed transitions between the levels form links (edges). It is feasible to characterize QM networks experimentally via high-resolution techniques of molecular spectroscopy. Thus, the term spectroscopic network (SN) was introduced for practical realizations of QM networks. The robust organizing principle of SNs is provided by QM selection rules; different transitions and transition intensities characterize different spectroscopic techniques. Even for the experimentally most thoroughly studied molecules the observable transitions form just a tiny part of all the allowed transitions. The complete list of allowed transitions can only be determined via sophisticated fourth-age quantum chemical computations.

Different intensity cutoffs for the transitions can be used to build first-principles SNs of different size. Figure 2 shows the visual representation of four such first-principles SNs of H\textsubscript{2}\textsuperscript{16}O, corresponding to 298 K one-photon absorption spectra with transition intensity cutoffs of 10\textsuperscript{-20}, 10\textsuperscript{-22}, 10\textsuperscript{-24}, and 10\textsuperscript{-26} cm molecule\textsuperscript{-1}. The data employed to generate Figure 2 are available in ref 42. The components corresponding to ortho- and para-H\textsubscript{2}\textsuperscript{16}O are clearly visible in Figure 2, as well as the buildup of vertices with a large number of edges, called hubs (section 2).

Viewing high-resolution molecular spectra as networks (Figure 3, vide infra) helps to answer a number of important as well as intriguing questions, some of which address shortcomings of the traditional, serial spectroscopic approach mentioned, including the following:

1. What would be the most economical way, i.e., the one based on the smallest number of feasible measurements, to determine the complete set of rovibronic states of a molecule?
2. What is the best way to validate existing spectroscopic measurements and to guide the design of future ones if efficiency is of prime concern?
3. Could one order the energy levels of a molecule based on their “importance”, and could we determine those transitions whose accurate knowledge is most important for improving the true accuracy of experimental line lists and related information systems?
4. How could experimental and theoretical high-resolution spectral data be unified and how could first-principles data be used to simplify the assignment of measured spectra?

The steps leading to answers to these questions require the understanding of the structure of SNs. For example, we need to know how many components would an experimental and a first-principles spectroscopic network possess. Then, we must...
understand whether the structure of an experimental SN is such that it supports only small components or it is characterized by a giant component or perhaps a few of them. We also need to know if any of these components, small or large, have any special structure; in other words, would the various selection rules of molecular spectroscopy dictate any special structure for the experimental and/or first-principles SNs? Furthermore, what kind of degree distribution would the links in the components of an experimental SN follow and would this be different from the distributions observable in the corresponding (truncated) first-principles SNs? How about not only first but also second and higher degree distributions?

It is certainly clear that interpretation and validation of the results of high-resolution spectroscopic experiments via network theory help to improve the accuracy, completeness, and robustness of line lists and guide related experiments. It is noted in this respect that having complete and accurate line lists for molecules of, for example, atmospheric and astrophysical interest at arbitrary temperatures is one of the “holy grails” of modern applied high-resolution spectroscopy. It is also important to emphasize the complementary nature of first-principles and measured line lists; although the relative accuracy of even the best first-principles energy levels is some 10−10000 times worse than that of typical experimental high-resolution transition data, most of the computed transition intensities have accuracies similar to those of experimental data. Thus, for the foreseeable future one needs to consider the combination of experimental and ab initio information to satisfy the needs of modellers.

The rest of the paper is organized as follows. Section 2 provides the graph-theoretical foundation necessary for the discussions in the subsequent sections of the article. Section 2 can be skipped by those familiar with at least the elementary ideas of modern network theory. Section 3 provides a detailed
account of the structure of SNs and how network theory can be used to better understand high-resolution spectroscopic results. Section 4 summarizes details about the presently most important application of SNs, the MARVEL (Measured Active Rotational–Vibrational Energy Levels) procedure, used to determine experimental-quality energy levels from all the available experimentally measured transitions. Section 5 summarizes and concludes our presentation.

2. GRAPH-THEORETICAL FUNDAMENTALS

Though some of the terms utilized are only defined later, we start the discussion with Figure 3, showing the connection between a measured high-resolution spectrum and its spectroscopic network in a simple, pictorial way.

The simulated spectrum shown in Figure 3 is the pure rotational absorption spectrum of H$_2^{16}$O corresponding to its ground vibrational state. Both the line positions and their intensities are taken from a first-principles exact nuclear-motion computation. Each line in the simulated spectrum is color-coded as follows: lines corresponding to ortho-H$_2$ are indicated with black and blue, whereas those of para-H$_2$ with green and red. For all transitions, the color of the lower half of the line corresponds to the lower state, whereas that of the upper half to the upper state. For the more intense lines a quantum-number assignment of the transition is also given, the numbers provided correspond to the rotational quantum number triad $J K_a K_c$ (vide infra). The transitions are indicated by arrows in the two graphs below the spectrum. There is a one-to-one correspondence between the two graphs containing (colored) energy levels and the spectrum containing (colored) transitions. The transitions are represented in the graphs as arrows, because in an absorption spectrum there is always a lower and an upper state and the arrows point from the lower to the upper state. Note that spectroscopic networks will be treated later as undirected graphs. As it is very clear from the two graphs, both for ortho- and para-H$_2$ the transitions always connect states where the $K_a$ values have different parity (odd vs even). This figure implies right away important facts about the spectrum of H$_2$; the graph corresponding to the spectrum contains two components (ortho- and para-H$_2$) and each component is a bipartite graph (this explains the use of two colors for each component). It is also evident that the graphs have only even-membered cycles, the smallest cycles have four vertices. The spectrum, at about 205 cm$^{-1}$, also shows an accidental degeneracy of two transitions, explaining the overlapping colors, one belonging to ortho- and one to para-H$_2$. The 3 times larger absorption intensity of the ortho transitions as compared to the corresponding para transitions, due to nuclear spin statistics, is also clearly visible in the figure, for example at about 150 cm$^{-1}$.

For the purposes of the present discussion, networks, and graphs are considered to be equivalent mathematical constructs. In what follows, in section 2.1, we give selected definitions of graph theory relevant for the study of spectroscopic networks. Some of the elementary definitions are shown graphically in Figure 4 to help those unfamiliar with graph theory. Section 2.2 introduces the most important matrix representations of networks. This is followed by important concepts useful to understand certain characteristics of SNs, including vertex ranking (section 2.3), and later we discuss complexity measures (section 2.4) and clustering techniques (section 2.5).

The most important terms useful to understand this paper are printed in bold in this section. Therefore, it is hoped that this section can serve as a glossary, where readers can turn to if becoming uncertain about the meaning of a particular term in the later parts of the paper.

2.1. Definitions. Intuitively, a graph is a representation of a set of objects where certain pairs of the objects (vertices) are connected by links (edges). Mathematically, a graph $G$ is an ordered pair, $G = (V, E)$, where $V$ is a set of vertices and $E$ is a set of edges, the edges being two-element subsets of $V$. If $V'$ is a subset of $V$, denoted by $V' \subseteq V$, and $E' \subseteq E$, then the graph $G' = (V', E')$ is a subgraph of $G = (V, E)$. We will write for the number of vertices $|V| = n$, and for the number of edges $|E| = m$.

Energy levels and transitions among the energy levels of a given quantum system are represented with the vertices and edges of the network $G$, respectively. An SN of a given molecule “mol” can be denoted as $G^{	ext{mol}}$. Further parameters characterizing a spectroscopic network $G$, e.g., temperature, measurement characteristics (for example, absorption or emission), and a transition intensity cutoff value, can be listed as subscripts to $G$. Note that the energy levels in an SN carry labels, usually made up of “good” and “approximate” quantum numbers. The labels have to be unique but occasionally it may be advantageous if they not only contain independent but also redundant information. A further peculiar characteristic of SNs is that the vertices, the rovibrionic energy levels, could be ordered on the basis of quantum numbers; perhaps the most useful quantum number is $J$, corresponding to the overall rotation of a molecule. This ordering would help placing states connected by strict selection rules close to each other.

The number of transitions (edges) that connect to an energy level (vertex) is called the degree of the energy level. Naturally, the sum of the energy level degrees is twice the number of transitions, $m$. Each edge $e = \{u, v\} \in E$ connects two adjacent vertices $u, v \in V$. If $u = v$, the edge is called a loop. SNs do not contain loops. Experimental (measured) SNs are usually multiedge graphs, containing multiple edges (between the
same pair of vertices), corresponding to multiple measurements of a certain transition. Graphs that contain neither loops nor multiple edges are called \textit{simple graphs}. First-principles SNs, in other words levels and transitions of the computed line list of a molecule corresponding to a chosen measurement technique, are simple graphs. If the edges have no direction, the graph is called \textit{undirected}; otherwise, it is called \textit{directed} (in a directed network the edges have a direction, pointing from one vertex to another). Spectroscopic networks are considered to be undirected and it is without loops and multiedges, usually containing more than one component. A component of a graph can have a designated vertex, called the \textit{root}. SNs may contain two types of roots, principal and other. \textbf{Principal roots} have a clear physical meaning (note here that the nodes of an SN can be ordered on the basis of their relative energies), as SNs often have two or even three principal roots based on nuclear spin isomerism.\footnote{The component of an SN containing a principal root is called a principal component (PC). A component without a principal root dictated by nuclear-spin isomerism is called a floating component (FC). FCs can also have roots but for them this concept does not appear to be useful. A straightforward and fast method for determining the components of a network is the Depth-First Search (DFS) algorithm.\cite{68} A giant component is a component whose size is of the same order as \textit{n}. If we index the vertices as \( V = \{v_1, \ldots, v_n\} \), then the \( i \)th row and column correspond to \( v_i \). Let \( d_i \) denote the degree of vertex \( v_i \). Then, the elements of the matrices mentioned are as follows:}

\begin{align}
A &= \begin{cases} 1, & \text{if there is an edge between } v_i \text{ and } v_j \\
0, & \text{otherwise} \end{cases} \\
L &= \begin{cases} d_i, & \text{if } i = j \\
-1, & \text{if there is an edge between } v_i \text{ and } v_j \\
0, & \text{otherwise} \end{cases} \\
L_N &= \begin{cases} 1, & \text{if } i = j \\
-(d_i d_j)^{-1/2}, & \text{if there is an edge between } v_i \text{ and } v_j \\
0, & \text{otherwise} \end{cases}
\end{align}

\footnote{Briefly, the adjacency matrix is used when the network study is aimed at adjacency relations, paths, and cycles. If the SN is undirected and it is without loops and multiedges, \textbf{A} is real and symmetric, with real eigenvalues and orthogonal eigenvectors. The symmetric combinatorial Laplacian matrix is a useful tool for dealing with the incidence relations of the edges and vertices and in the matrix tree theorem.\footnote{Both Laplacian matrices are popular for vertex clustering methods.\cite{73,74} See section 3.9 for some of the uses of \textbf{A}, \( L^C \), and \( L^N \) in the case of SNs. Finally, in ref 54 we introduced the Ritz-matrix of spectroscopy, to honor the contributions of Walther Ritz to the field. The Ritz-matrix \( X \) of an SN is of size \( n \times m \), and it is similar to the incidence matrix of a graph in the directed sense:}

\begin{align}
X &= \begin{cases} +1, & \text{if } i \text{ is the upper energy level of transition } \epsilon_j \\
-1, & \text{if } i \text{ is the lower energy level of transition } \epsilon_j \\
0, & \text{otherwise} \end{cases}
\end{align}
The Ritz-matrix plays a vital role in the X-matrix technique introduced by Flaud et al., forming the basis of the MARVEL procedure, at present the principal application based on the idea of SNs (section 4).

2.3. Vertex Ranking. The simplest approach to quantify the importance, called “centrality” in some applications of network theory, of vertices within a set V of an SN is to use the degree distribution. In this case, the most important node will be the energy level with the largest number of associated transitions, the second with the second largest degree, and so on. See sections 3.3 and 3.4 for the degree distributions of experimental and first-principles SNs.

A complementary approach, employed successfully in search optimization engines, is the use of the PageRank, a network-based diffusion algorithm. A central feature of this algorithm is that importance depends not only on the number of incoming links but also on the “quality” of the links: a link coming from an important node is much more valuable than a link from an insignificant one. In SNs, this translates to the requirement that regarding the relative importance of an energy level a link to an important energy level, itself having a large requirement that regarding the relative importance of an energy level a link to an important energy level, itself having a large requirement.

The recursive form of the PageRank determination is as follows:

\[
PR(i) = \frac{1 - \alpha}{n} + \alpha \sum_{j \in M(i)} \frac{PR(j)}{d_j}
\]

where \(PR(i)\) is the PageRank of the \(i\)th node, \(M(i)\) is the set of neighbors of the \(i\)th node, and \(\alpha \in (0, 1)\) is a dampening factor. It is important to note that if we take an ordering of the vertices based on their PageRank value, changing \(\alpha\) does not alter the order.

The Pearson correlation coefficient of the degrees at either end of an edge is defined as

\[
r(G) = \frac{\sum_{i,j \in T} d_i d_j - \left(\sum_{i,j \in T} d_i + d_j\right)^2}{\sum_{i,j \in T} d_i^2 + d_j^2 - \left(\sum_{i,j \in T} d_i + d_j\right)^2}
\]

where \(l\) is the number of edges in the graph. \(r(G)\) has been introduced to analyze the assortativity of a network. Social networks, for example, usually show assortativity mixing on their degrees, that is, their high-degree vertices like to connect to other high-degree vertices in the network. The opposite of assortativity mixing is disassortativity mixing, whereby high-degree vertices attach to low-degree ones. The latter property is frequent in technological and biological networks.

2.5. Vertex Clustering. The principal aim of vertex clustering is to distribute the vertices of a network into clusters along predefined properties. The subsets the partitioning yields must be pairwise disjoint, and their union should be the original vertex set. If we distribute a partition into \(k\) subsets, we call it a \(k\)-clustering.

Clustering algorithms can be divided into at least two classes: partition and hierarchical clustering. Partition clustering methods generate a starting partition first, then re-evaluate the actual clustering in each step in an iterative loop until arrive at the required clustering result. Hierarchical clustering algorithms are using the iteration of either merging smaller clusters to larger ones (in agglomerative clustering), or dividing larger clusters into smaller ones (in divisive clustering), based on a similarity measure. Both types of clustering result in a hierarchy of clusters, called a dendrogram.

Generally, partition clustering techniques are faster, but they require more information beforehand (for example, the number of clusters in the network). The Clauset-Newman-Moore (CNM) algorithm is an efficient hierarchical agglomerative algorithm that can be used for clustering of even large networks, like SNs. It is useful for highlighting communities within a network. One can use the Stanford Network Analysis Platform (SNAP) for CNM clustering. Maintaining a balance in the size of the subsets during the clustering is definitely a preferred property. Normalization means that the clusters obtained should have roughly the same size.

Spectral clustering algorithms are applicable for partition clustering. To determine a partitioning of the energy levels of one of the PCs of an experimental SN, the spectral clustering technique utilizes the eigenvectors of the normalized Laplacian matrix, \(L^N\), of the network. It is especially meaningful in spectroscopy to look for a partitioning where the number of edges is minimal between the different clusters. This way one forms “communities” within the network: a community \(G^\prime\) is defined when the sum of all degrees of vertices within \(G^\prime\) is (much) larger than the sum of all degrees toward the rest of the network. To obtain a \(k\)-clustering of this type, one should determine \(k\) eigenvectors corresponding to the largest eigenvalues of \(L^N\), form a matrix of size \(n \times k\) where the eigenvectors are in the columns, normalize the rows, and then apply a \(k\)-means clustering to the rows, which are considered to be \(n\) \(k\)-dimensional points.

Another useful concept that can be employed in high-resolution molecular spectroscopy is that of the bridge.

\[
s = \sum_{i,j \in E} d_i d_j
\]

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

\[
s_{\text{max}} = \sum_{i=1}^{n} \frac{d_i^3}{2}
\]

we can define the normalized structural metric, \(S(G)\), as

\[
S(G) = \frac{s}{s_{\text{max}}}
\]
bridge is an edge of a network whose deletion increases the number of connected components. Naturally, a bridge may not be a member of any cycle. One of the simplest (and fastest) bridge-finding techniques uses the DFS algorithm. Finding bridges in SNs is useful because (a) exploring connectors of larger subgraphs can help to explore the weakly connected components of the SN (if an energy level is a member of a weakly connected cluster, then its uncertainty highly depends on the accuracy of the bridge) and (b) exploring branches (Figure 9, vide infra) helps to detect those energy levels whose values are the least reliable.

3. STRUCTURE OF SPECTROSCOPIC NETWORKS

Based on the network definitions and concepts presented in section 2, the questions raised toward the end of the Introduction can be answered and a number of useful statements can be developed about different aspects of the structure of SNs. These statements contribute to our improved understanding of the results of high-resolution molecular spectroscopic experiments as well as of data deposited in spectroscopic information systems.

For some of the structural properties it is important to distinguish between experimental and first-principles SNs of a molecule. Construction of an experimental SN is obvious: one needs to collect all assigned transitions from the literature (see Table 1 for the number of sources identified during the building of SNs for the molecules studied thus far). Construction of a first-principles SN goes through the following steps: (a) take all (available) computed energy levels as nodes; (b) use the selection rules appropriate for the molecule and the experiment to link the nodes; and (c) add the computed intensities, serving as network weights, as well, to the links based on the type of experiment and the chosen temperature. Transition intensity cutoff values can be used to select a subset of the possible links, as done in the panels of Figure 2. In this section we usually concentrate on experimental SNs, whereas first-principles ones receive somewhat less attention.

So far, the experimental SNs of 15 molecules have been investigated. The list of molecules include nine isotopologues of water, three isotopologues of HCl,13,15 three isotopologues of HCN,12,55 and ketene, CH3C=O.56 Table 1 lists some of the principal characteristics of the experimental SNs of these molecules. Table 2 lists data relevant for spectroscopic networks contained in the canonical line-by-line information system HITRAN and investigated in this study. The HITRAN database contains not only experimental data but also accurate computed ones when available and needed. A note about the HITRAN data used in this study: among the large number of transitions present in HITRAN for a large number of molecules, there are some that seemingly correspond to forbidden transitions. For example, the HITRAN data for H3O+ contain five transitions that violate the ortho–para selection rule. The frequencies of these transitions are 11069.73629, 19430.03950, 19706.8374, 19870.5948, and 21295.3901 cm⁻¹. These transitions were removed from the present analysis. Another example is the H2O molecule, where the HITRAN data lead to the presence of a three-membered cycle. This odd-membered cycle violates the one-photon selection rules and the bipartiteness of the corresponding SN (subsection 3.2, vide infra); therefore, these transitions, at 18.975258, 1387.383507, and 1393.928115 cm⁻¹, were also removed before our analysis. Most of the discussion within this section focuses on data in Tables 1 and 2.

3.1. Components of SNs. First-principles complete SNs of molecules may have several principal components (PCs, not to be confused with the unrelated principal components analysis), as required by nuclear spin statistics.44 Each PC has a root, which can conveniently be chosen as the lowest-energy level of the PC. PCs are expected to be giant components of SNs. Selection rules put constraints on the changes of the ("good" and "approximate") quantum numbers describing the energy levels involved in the transitions within a SN. Under field-free conditions, transitions are not allowed among the energy levels belonging to different PCs.44 Therefore, it is challenging to measure the energy difference between the roots of the PCs (vide infra).

As an example, we note that symmetric isotopologues of water (e.g., H2O, D2O, T2O, or the similar three isotopologues of D2O) have two principal components, traditionally44 called ortho and para [depending on whether the spins of the protons are parallel (ortho, total nuclear spin I = 1) or antiparallel (para, I = 0)]. On the contrary, the SN of the HD16O isotopologue of water, with all of its nuclei different, has only a single PC, i.e., all of the energy levels of HD16O form part of the same first-principles SN.

Figure 5 shows the lowest-energy part of the complete purely rotational one-photon absorption SN of the HD16O molecule, up to J = 3, where J is the rotational quantum number and Kc and Ke are the usual asymmetric-top quantum numbers.44,45 The selection rules behind the connectivity of the energy levels state that if ΔJ = 0(±1) then p = ±1(0), where p is the parity describing the energy level. These rules dictate that if the rotational parity, defined as (−1)Kc, is even or odd, then the maximum number of pure rotational transitions from a given rovibrational energy level with asymmetric-top label46 J_{K_c,K_e} is 3J + 2 or 3J + 1, respectively, as shown in Figure 4. It is important to observe that not all rotational transitions emanating from even the lowest-J states have been measured, though the HD16O molecule is thoroughly studied experimentally; see ref 41 and the data in Table 1. As J increases, the number of transitions not measured increases rapidly (not shown in Figure 5). Note that although many of the transitions may not be measured directly, their absence does not hinder obtaining the complete tree structure of the SN. Furthermore, identification of the missing transitions in experimental spectra becomes straightforward if the "inverted" rovibrational energies are available and the transition intensity, perhaps computed ab initio, is sufficiently large.47 Note also the apparent lack of odd-numbered cycles in Figure 5.

<table>
<thead>
<tr>
<th>HITRAN index</th>
<th>name</th>
<th>no. of nodes</th>
<th>no. of unique links</th>
<th>no. of components</th>
<th>bipartite</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>water</td>
<td>17045</td>
<td>134063</td>
<td>2</td>
<td>true</td>
</tr>
<tr>
<td>3</td>
<td>ozone</td>
<td>38816</td>
<td>260094</td>
<td>24</td>
<td>true</td>
</tr>
<tr>
<td>9</td>
<td>sulfur dioxide</td>
<td>18054</td>
<td>72460</td>
<td>1</td>
<td>true</td>
</tr>
<tr>
<td>10</td>
<td>nitrogen dioxide</td>
<td>6931</td>
<td>26334</td>
<td>5</td>
<td>true</td>
</tr>
<tr>
<td>20</td>
<td>formaldehyde</td>
<td>9847</td>
<td>40670</td>
<td>14</td>
<td>true</td>
</tr>
<tr>
<td>25</td>
<td>hydrogen peroxide</td>
<td>14975</td>
<td>126949</td>
<td>24</td>
<td>true</td>
</tr>
<tr>
<td>32</td>
<td>formic acid</td>
<td>11385</td>
<td>62684</td>
<td>1</td>
<td>true</td>
</tr>
</tbody>
</table>
Experimental SNs form multiedge graphs. As measurements done by different groups use different techniques and spectrometers working in different regions of the electromagnetic spectrum, there is some randomness in how transitions become measured. Thus, in principle spectroscopic measurements could result in several components. We did not find a case where the experimental SN of a molecule contained a large number of FCs, in all cases the great majority of the energy levels and transitions were part of the PCs of the network.

The size of a first-principles SN, both in the number of nodes and especially in the number of links, depends heavily on the chosen transition intensity cutoff. We did not find a case where first-principles SNs contained at whatever intensity cutoff value, floating components, the transitions selected this way always belonged to the PCs. Thus, both the experimental and the first-principles SNs seem to have giant components, an observation that can be explained by the degree distribution of SNs (see subsections 3.3 and 3.4).

Attempts to unite the components of an experimental SN are important for several reasons, most importantly as they (a) allow the attachment of proper, “absolute” energy values to the vertices of the components disjoint from the PCs having the energy zero; (b) help to improve the robustness of the SN; and (c) may lead to the design of new experiments and/or suggest to study certain spectral regions to improve the information content of the SN.

As shown in ref 54, even for the smallest experimental SNs studied there is an extremely large number of possible spanning trees for each of the PCs, on the order of $10^{10}$–$10^{50}$. Nevertheless, because a unique weight can be assigned to each link (a unique transition intensity in the cases studied), there will be a unique minimum-weight spanning tree for each component. Investigation and comparison of the structure of the experimental and first-principles minimum-weight spanning forests provides the simplest and most efficient way to connect the possible FCs to PCs, the giant components of the SN.

Finally, there is another statement that the investigation of the structure of first-principles SNs of different size yields:

---

**Figure 5.** Experimental purely rotational one-photon absorption spectroscopic network of HD^{16}O in the ground vibrational state up to $J = 3$ (the root of the graph is the $J_{KK} = 0_00$ state, the notation is changed in the figure to $J_{KcKc}$). The number in parentheses in each oval, representing an energy level, provides the number of allowed rotational transitions starting or ending on the given energy level within the ground vibrational state. The bipartite nature of the network is indicated with the two different background colors of the energy levels. The nodes (energy levels) are connected if the transition has been measured experimentally.

**Figure 6.** Pictorial representations and degree distributions of random networks of Erdős–Rényi (left panels) and of scale-free (right panels) character, the so-called hubs (nodes with a large number of links) are shown in orange in the latter case.
although the great majority of the transitions is extremely weak, it is easy to find a few relatively strong transitions for almost all vertices in the SN. This suggests that with relatively standard spectroscopic measurements almost all of the energy levels can be determined, which in turn could lead to an almost complete knowledge of high-resolution experimental absorption (and emission) spectra if transitions based on “measured” energy levels were augmented with computed line intensity values.

3.2. Bipartiteness. There is a highly special and highly restrictive property of SNs related to the quantum mechanical selection rules: all the SNs investigated thus far (see the entries of Tables 1 and 2 for the list of molecules) are bipartite. This reflects an important property of rovibrational states, namely their overall parity, which has to change during the selection rule of molecular rovibrational transitions is not violated, there are no simple cycles of odd length in SNs, the smallest cycle must involve four energy levels and there are only even-numbered cycles in SNs.

As shown in Table 2, all the SNs selected from the canonical spectroscopic database, HITRAN,25 are bipartite. This suggests that the spectroscopic data in HITRAN are correct for these molecules, at least in this sense.

3.3. Degree Distribution of SNs. In network theory a network G with n vertices and m edges is said to be sparse if m \( \ll n^2 \) and dense if m = \( O(n^2) \). Clearly, all experimental and first-principles SNs are sparse networks.

Erdős–Rényi random graphs are built by a process whereby links are placed randomly between a fixed number of vertices. Erdős–Rényi random graphs with a fixed set of \( V_{\text{fix}} \) vertices have a “characteristic” degree \( 2E/V_{\text{fix}} \), i.e., the vertex degrees have approximately a Poisson distribution with a mean of \( 2E/V_{\text{fix}} \) (Figure 6). If a random graph is allowed to grow and links are added on the basis of probabilities proportional to the momentary degrees of the vertices, a scale-free (SF) random graph results. The scale-free property of a network means that the probability that a randomly selected node has exactly d links is \( P(d) \propto d^{-\gamma} \), where \( \gamma \) is called the scaling index. SF random networks, as opposed to Erdős–Rényi random networks, are characterized by (a) a relatively few nodes with a large number of links (these nodes are called hubs), and (b) a robust connectivity structure hard to fragment by random removal of nodes. The following dynamical features are the usually assumed requirements of a SF random network: (a) evolutionary growth with more or less random generation of new nodes, (b) highly interactive self-organization, and (c) preferential connectivity of new nodes to old ones. Hubs shorten the paths between vertices of a network.

Features of SF networks mentioned occur naturally for SNs on the basis of spectroscopic experiments. For complete first-principles SNs this is not true; nevertheless, if the intensities of the transitions, always spanning many orders of magnitude, are taken into account, first-principles SNs also become scale free. Naturally, the size of a first-principles SN, in the number of both energy levels (nodes) and transitions (links), depends on the chosen cutoff of the absorption intensities. The computed

Figure 7. Size–frequency \([k–P(k)]\) plots for measured (panel A) and first-principles computed (panel B) transitions for HD\(^{16}\)O, with an absorption intensity cutoff of \( 10^{-22} \) cm molecule\(^{-1}\) in the latter case.

Table 3. Power-Law Distribution Fittings to the Principal Components (pc) of the Spectroscopic Networks of Selected Molecules Present in the HITRAN Information System

<table>
<thead>
<tr>
<th>HITRAN index</th>
<th>name</th>
<th>no. of nodes</th>
<th>no. of links</th>
<th>( \gamma )</th>
<th>( x_{\text{min}} )</th>
<th>( x_{\text{max}} )</th>
<th>KS p</th>
<th>avg degree</th>
<th>diam</th>
<th>avg dist</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>water pc#1</td>
<td>7871</td>
<td>59560</td>
<td>2.01</td>
<td>150</td>
<td>805</td>
<td>0.39</td>
<td>15.1</td>
<td>27</td>
<td>5.8</td>
</tr>
<tr>
<td>1</td>
<td>water pc#2</td>
<td>9174</td>
<td>74503</td>
<td>2.05</td>
<td>180</td>
<td>879</td>
<td>0.27</td>
<td>16.2</td>
<td>28</td>
<td>6.0</td>
</tr>
<tr>
<td>3</td>
<td>ozone</td>
<td>38719</td>
<td>260006</td>
<td>2.26</td>
<td>667</td>
<td>5666</td>
<td>0.95</td>
<td>13.4</td>
<td>88</td>
<td>20.0</td>
</tr>
<tr>
<td>9</td>
<td>sulfur dioxide</td>
<td>18054</td>
<td>72460</td>
<td>1.99</td>
<td>54</td>
<td>5919</td>
<td>0.87</td>
<td>8.0</td>
<td>101</td>
<td>25.8</td>
</tr>
<tr>
<td>10</td>
<td>nitrogen dioxide</td>
<td>6778</td>
<td>26086</td>
<td>1.76</td>
<td>31</td>
<td>1870</td>
<td>0.99</td>
<td>7.7</td>
<td>103</td>
<td>22.4</td>
</tr>
<tr>
<td>20</td>
<td>formaldehyde pc#1</td>
<td>5007</td>
<td>21656</td>
<td>2.18</td>
<td>200</td>
<td>795</td>
<td>0.41</td>
<td>8.7</td>
<td>65</td>
<td>13.7</td>
</tr>
<tr>
<td>20</td>
<td>formaldehyde pc#2</td>
<td>4610</td>
<td>18796</td>
<td>2.84</td>
<td>300</td>
<td>856</td>
<td>0.33</td>
<td>8.2</td>
<td>62</td>
<td>13.7</td>
</tr>
<tr>
<td>25</td>
<td>hydrogen peroxide pc#1</td>
<td>7604</td>
<td>69160</td>
<td>2.23</td>
<td>51</td>
<td>469</td>
<td>0.87</td>
<td>18.2</td>
<td>50</td>
<td>14.2</td>
</tr>
<tr>
<td>25</td>
<td>hydrogen peroxide pc#2</td>
<td>7213</td>
<td>57650</td>
<td>2.52</td>
<td>86</td>
<td>723</td>
<td>1.00</td>
<td>16.0</td>
<td>50</td>
<td>14.3</td>
</tr>
<tr>
<td>32</td>
<td>formic acid</td>
<td>11385</td>
<td>62684</td>
<td>2.15</td>
<td>200</td>
<td>1610</td>
<td>0.38</td>
<td>11.0</td>
<td>81</td>
<td>22.9</td>
</tr>
</tbody>
</table>

*PC = principal component, diameter = diameter, avg = average, dist = distance. See the text for further definitions employed in the table.
SN of HD\textsuperscript{18}O contains\textsuperscript{16} altogether 163491 energy levels and 69744828 transitions. At 298 K, if the absorption intensity cutoff is chosen to be 10\textsuperscript{-20} cm molecule\textsuperscript{-1}, the SN contains 73590 nodes and 863575 links. With still realistic cutoffs of 10\textsuperscript{-20} and 10\textsuperscript{-28} cm molecule\textsuperscript{-1}, the first-principles SN of HD\textsuperscript{18}O contains 128106 (4720711) and 153600 (15356682) nodes (links), respectively. From the numbers presented it is clear that the number of energy levels grows much more slowly than the number of transitions, the great majority of the transitions is extremely weak. All these SNs are characterized by a SF distribution; see two representative examples in Figure 7.

The degree distributions of the molecules selected from the HITRAN database, Table 3, also exhibit heavy tails. The \( \gamma \) column of Table 3 contains scaling indices for 10 principal components. The \( x_{\text{min}} \) values are the lower bounds for the vertex degree frequencies considered during the fitting. KS\( p \) is the \( p \)-value of the Kolmogorov–Smirnov test with the corresponding \( \gamma \) and \( x_{\text{min}} \) parameters. The \( x_{\text{max}} \) parameter denotes the highest among the degree frequencies in the component. It seems that for most molecular species, if not for all, the degree distribution is scale-free. Some of the most important consequences of this observation are discussed in section 3.5. Furthermore, the scaling indices are all larger than 2.0; their average value is about 2.2.

### 3.4. First- and Second-Degree Distributions

There is no established spectroscopic law requiring that the most intense lines, those that can be most easily measured, should be those that can be most easily measured. As shown in Table 1 of ref\textsuperscript{52}, the most relevant statements are as follows: (a) a network will have a giant component if \( \gamma < 3.4788... \) but not if it is larger; (b) the giant component corresponds to the entire network when \( \gamma \geq 2 \); and (c) in the region between \( \gamma = 2 \) and \( \gamma = 3.4788... \) there is a giant component but it does not fill the whole network. Because experimental SNs seem to have \( \gamma \approx 2.2 \), Table 2, these modeling studies nicely support the observation that experimental SNs have giant components and occasionally some small ones. Nevertheless, note that first-principles SNs obtained with reasonable intensity cutoff criteria seem to have only giant components.

The degree distribution completely determines the statistical properties of uncorrelated networks. As the related investigations show,\textsuperscript{17} a large number of real networks are correlated, meaning that the probability that a node of degree \( k \) connects to a node of degree \( k' \) depends on \( k \).

### 3.5. Hubs and Edge Densities

Scale-free networks contain a small number of hubs. As expected, the hubs with the largest number of degrees in a one-photon absorption SN are on the ground vibrational state. For the measured SN\textsuperscript{34} of G\textsuperscript{18}O\textsuperscript{2}\ the numbers of the largest number of connections are \( 6_{34}(4042) \), \( 7_{35}(3970) \), and \( 6_{56}(3897) \), where the number of links is given in parentheses.

The scale-free property of SNs means that despite the fact that SNs can be extremely large (though always finite in realistic cases), there are only relatively few energy levels whose accuracy principally determines the overall accuracy of the energy levels of an experimental SN. Experiments which improve the accuracy of SNs by decreasing the uncertainties of energy levels qualifying to be hubs are the most useful ones. This leads to the important conclusion that all microwave (MW), millimeter wave (MMW), and far-infrared (FIR) measurements performed with this aim in mind would be highly beneficial for improving the overall accuracy of experimental SNs of small molecules of prime interest for sophisticated modeling studies.

For undirected simple graphs, like the SNs, the edge density is defined as \( D = 2\langle E \rangle / [V(V-1)] \). Although in simple graphs the maximum number of edges is \( V(V-1)/2 \), in SNs this is much smaller due to the constraints provided by the quantum mechanical selection rules. As shown in Table 1 of ref \textsuperscript{52}, the

### Table 4. First and Second Moments, and the Mean Number of the Second and Third Neighbors, in the Experimental Spectroscopic Networks of Selected Small Molecules

<table>
<thead>
<tr>
<th>molecule</th>
<th>ref</th>
<th>( \langle d \rangle )</th>
<th>( \langle d^2 \rangle )</th>
<th>( c_2 )</th>
<th>( c_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma \cdot \text{H}_2\text{O} )</td>
<td>42</td>
<td>11.89</td>
<td>839.26</td>
<td>827.37</td>
<td>57572.8</td>
</tr>
<tr>
<td>( \rho \cdot \text{H}_2\text{O} )</td>
<td>42</td>
<td>9.05</td>
<td>633.07</td>
<td>624.02</td>
<td>43027.7</td>
</tr>
<tr>
<td>( \sigma \cdot \text{NH}_3 )</td>
<td>58</td>
<td>11.50</td>
<td>683.71</td>
<td>672.21</td>
<td>39329.7</td>
</tr>
<tr>
<td>( \rho \cdot \text{NH}_3 )</td>
<td>58</td>
<td>11.05</td>
<td>629.94</td>
<td>618.89</td>
<td>34662.7</td>
</tr>
<tr>
<td>( ^{12}\text{C}_2 )</td>
<td>59</td>
<td>6.16</td>
<td>69.87</td>
<td>63.71</td>
<td>659.1</td>
</tr>
</tbody>
</table>

\[
 c_2 = \frac{c_1}{c_1} d^{-1}
\]  

(12)

For comparative purposes, Table 4 contains the values of \( c_3 \) obtained this way.
edge density $D$ of the first-principles SNs of HD$^{16}$O, with intensity cutoff values ranging from $10^{-20}$ to $10^{-90}$ cm molecule$^{-1}$, has a minimum at about $10^{-50}$ cm molecule$^{-1}$. This minimum is achieved at an intensity value that is about the limit for present-day absorption measurements. Nevertheless, it must not be forgotten that the investigated line list becomes incomplete at about this intensity cutoff value.

As shown in ref 53, the behavior of hubs within the SN can be described using graph metrics (section 2.4). For the first-principles model of the H$_2^{16}$O molecule with an intensity cutoff $10^{-20}$ cm molecule$^{-1}$, we have an $r(G)$ value that is close to zero, with a relatively large $S(G)$. This translates to the fact that hubs like to connect to each other in the spectroscopic network, although the hubs also have many low-degree neighbors. Moreover, as we reduce the absorption intensity cutoff parameter ($10^{-22}$, $10^{-24}$, $10^{-26}$, and $10^{-28}$ cm molecule$^{-1}$), it becomes obvious that the SN becomes increasingly disassortative, as many new transitions appear with low-degree nodes at one or both end points. This in turn lowers the ratio of the edges which connect two high-degree nodes in the SN.

### 3.6. PageRank

Determining a useful importance ordering (“centrality measure”) for the set of energy levels of a SN is important for orienting high-resolution spectroscopists and spectroscopic information system developers, as well. Importance is an intuitive term here, but it refers to various enquiries, including (1) which energy levels have the largest number of transitions associated with them, (2) which energy levels are present in the largest number of cycles in the SN, or (3) which energy levels are affecting the uncertainties and their propagation in the SN the most.

Question 1, for example, can be answered by setting up the $A$ matrix, as the degree $d_i$ of vertex $i$ of $G$, i.e., the number of its connections, is simply $d_i = \sum_n A_{ni}$. Note that the degree of an energy level may not reflect properly the importance of even a hub, as most connections may be made to low-degree energy levels. Answering question 2 is a difficult and computationally demanding task, especially because SNs have a very large number of cycles of very different size (see Figure 3 for a simple case). Nevertheless, answering the question whether the energy level is part of even one cycle, which should significantly help its accurate determination, is straightforward. The answer to question 3 will be discussed in some more detail in the subsection on network clustering.

The PageRank order of the hubs can be significantly different from their degree order. PageRank ordering appears to be a more useful measure to judge the relative importance of the most important energy levels (hubs) as within the PageRank ordering hubs are preferentially connected with hubs. For H$_2^{16}$O, for example, the most important hubs with the highest degrees and the highest PageRanks are all on the vibrational ground state and have $J$ values around 5.

An example to demonstrate the usefulness of the PageRank ordering was given in ref 54 during the investigation of the SN of ortho-H$_3^+$ with respect to degree ordering (left panel) and PageRank ordering (right panel). Notice that only four nodes, those colored green in the figure, are present in both lists. Moreover, the subgraph defined by the ten nodes having the highest degrees is disconnected, in other words, these nodes do not form one connected component, instead, they form six. When PageRank defines the centrality, the subgraph is well connected with a lot of cycles, with an eminent participation of hubs in these cycles. As cycles are especially important in validating measurements, we must consider the use of the PageRank ordering advantageous.

### 3.7. Diameter

There are additional signatures beyond the power-law degree distribution that characterize SF networks. An important one is the small-world property: scale-free networks are characterized by a small diameter.

The diameter of a network is defined as the maximum distance of a vertex pair; in other words, the longest path of the shortest paths among all vertex pairs. The diameter of the SN can be estimated by counting the different eigenvalues of $A$: if $A$ has $r$ different eigenvalues, then the diameter of $G$ is at most $r − 1$.

The interconnectedness of a particular SN can be described efficiently by the diameter. The diameter computed statistically for the measured SN of HD$^{16}$O is only about 7, though it is still considerably larger than the corresponding $log v$ value, about 3. This value is slightly larger but similar to the diameter of the first-principles SN. As the absorption intensity cutoff is decreased, the diameter of the computed SN seems to stay around this value though it becomes somewhat smaller. Thus, SNs clearly have an intrinsic small-world property, similarly to most other complex networks studied in nature, society, communication, and elsewhere.

### 3.8. Network Vulnerability

Selection rules allow only a limited number of links between the nodes of the SN. As the SN becomes larger, either via new measurements for an experimental SN or by a decrease in the intensity cutoff for a first-principles SN, the number of links increases substantially but not the number of nodes. The number of (even-membered) cycles within the network also increases dramat-
Robustness of SNs can be ascertained by a numerical experiment involving random removal of nodes.\textsuperscript{5,52} The results for three selected SNs corresponding to HD\textsuperscript{16}O, one being the measured SN, the other a first-principles SN, and the third the purely rotational first-principles SN, are shown in Figure 9.

As shown in ref\textsuperscript{54}, this formula gives extremely large values for the number of spanning trees even for relatively small SNs. Thus, weighting introduced to SNs serves a special purpose if spanning trees are to be used for high-resolution spectroscopy.

### 3.10. Clustering

Connectors between relatively dense subnetworks (clusters) of SNs can be identified and analyzed via several variants of spectral clustering techniques\textsuperscript{73} based on the combinatorial and normalized Laplacian matrices introduced in section 2, \( L^C \) and \( L^N \), respectively.

The partition and hierarchical clustering results appear to have considerable value in identifying the “weakest links”, from an information system point of view the most significant links in the SN. Identification of these links is especially important as they may limit the accuracy of the determination of a large number of energy levels, separated from well-defined energy levels by the small number of connectors, even if these energy levels form part of several (local) cycles. The identification of “weakest links” becomes especially important when one judges the true accuracy of the experimental rovibrational energy levels obtained through the MARVEL approach, which converts information in measured transitions to information about the energy levels of a molecule.

The number of bridges in the experimental SNs of H\textsubscript{2}O, \( ^{14}\text{NH}_3 \), and \( ^{12}\text{C}_2 \) of the present analysis is 3513, 730, and 1296, respectively. Comparing these values with the total number of links (98838, 15393, and 16481, respectively) suggests that, in a relative sense, it is the database of \( ^{12}\text{C}_2 \) that contains the largest number of bridges. This explains why the diameter of the experimental SN of \( ^{12}\text{C}_2 \) is much larger than usual (see Table 2 of ref\textsuperscript{54}). The number of energy levels whose values depend on bridge transitions are 4865, 965, and 1575 for H\textsubscript{2}O, \( ^{14}\text{NH}_3 \), and \( ^{12}\text{C}_2 \), respectively. The energy level values might be incorrect because they rely on the correctness and accuracy of a single transition.

### 3.11. Data Reduction via SNs

Because high-resolution spectroscopic measurements yield an extreme amount of information, the reduction of the observed data to manageable size is a basic challenge for the theory of spectroscopy. The standard solution is to use model Hamiltonians with a relatively small number of parameters and least-squares optimize these parameters to represent all the measured data.\textsuperscript{90} In a way this means that spectroscopic transitions are converted to parameters yielding energy levels (and transitions via the Ritz principle). These parameters allow excellent interpolation but they may fail when used to extrapolate beyond the measured
range (especially if one considers the extremely high accuracy of most of the measurements).

SNs offer another, somewhat less spectacular data reduction facility via the inversion of transitions to energy levels. The best way to reduce the information content of experimental transitions is through the use of weighted spanning trees corresponding to the experimental SN. This way one can reduce the information contained in the huge number of measured transitions of the network to a relatively small set of energy levels and transitions (the saving is about an order of magnitude, Table 1, but grows fast as the size of the experimental SN grows). Nevertheless, to judge the true accuracy of the energy levels, the information contained in a tree (forest) is not sufficient as it is only through cycles that one can determine the true accuracy of the measurements (section 4). The network-theoretical view allows us to appreciate how (even-membered) 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 could tighten their uncertainties even below those of the original measurements.

3.12. Assigning Spectra. Assigning complicated high-resolution spectra is a major challenge; consequently, high-resolution spectroscopy is also a science (and art) of the quantum number assignment of measured lines. The techniques used evolved much over the years, but the end results is about the same: a high-resolution spectrum of a polyatomic molecule is converted to a list of labeled energies (Figure 3). When spectroscopists analyze high-resolution experimental spectra, they traditionally associate the lines with some good and mostly approximate quantum numbers by mapping the observed spectrum and the subsequent analysis time exceeds an acceptable limit.

In ref 53 we advocated a novel protocol for the assignment of high-resolution one-photon absorption spectra based on the concept of SNs: detect the lines in a measured high-resolution spectrum leading to the largest number of new energy levels via a small number of spectroscopic parameters of a well-designed model Hamiltonian. This type of assignment procedure fails in the case of spectroscopic parameters of a well-designed model Hamiltonian. This type of assignment procedure fails in the case of spectroscopic parameters of a well-designed model Hamiltonian. This type of assignment procedure fails in the case of spectroscopic parameters of a well-designed model Hamiltonian.

The MARVEL approach is designed for a critical evaluation and consequent validation of experimental transition wavenumbers and uncertainties collected from the literature (Table 1), followed by the inversion of the wavenumber information to obtain the best possible set of energy levels with attached dependable uncertainties. Briefly, the MARVEL protocol includes the following steps: (1) Collection, preliminary validation, and compilation of all the available measured transitions possessing unambiguous labels and uncertainties into a database. (2) Determination of the distinct energy levels of the SN, built from the measured data collected. (3) Setting up a \( N_l \)-dimensional vector, \( Y \), containing the experimentally measured transitions, an \( (N_l - 1) \)-dimensional one, \( X \), containing the energy levels sought, and an extremely sparse matrix, \( a \), of dimension \( N_l \times (N_l - 1) \), the Ritz-matrix, describing the relation between the transitions and the energy levels. (4) Least-squares solution of the system of linear equations obtained, including iterative improvement\(^{99,100}\) of the experimental uncertainties if needed. In all these steps highly efficient algorithms are needed due to the large size of the SNs. These algorithms have been found; thus, one iterative step in the MARVEL process, involving the formal inversion of a \( 100000 \times 100000 \) matrix, takes less than a second on a single core.

The fundamental equations behind MARVEL are extremely simple. First, MARVEL is built upon the Ritz principle, which gives the connection between the measured transitions and the rovibronic energy levels:

\[
\sigma_{ij} = E_i - E_j
\]  

(14)

where \( \sigma_{ij} \) is a measured wavenumber \( (ij = 1, ..., N_l) \), and \( E_i \) is a lower and \( E_j \) is an upper rovibronic energy level \( (i, j \in 1, ..., N_l) \). Let \( \delta_{ij} \) be the measurement uncertainty of the \( \sigma_{ij} \) transition. Thus, the principal input to MARVEL is a grand list of \( N_l \) experimentally measured, assigned, and labeled transitions with corresponding uncertainties, and the aim of MARVEL is to determine the \( N_l \) energy levels with self-consistent uncertainties. An overdetermined system of linear equations,

\[
aX = Y
\]  

(15)

characterizes all SNs. When weights \( w_{ij} = \delta_{ij}^{-2} \) are introduced, we can write

\[
aX = B
\]  

(16)

where \( A = a^Twa \) and \( B = a^Twy \), and the dimension of the extremely sparse \( A \) matrix is \( (N_l - 1) \times (N_l - 1) \). Computation of \( A \) and \( B \) can be considerably accelerated by the use of analytic formulas.\(^{50}\) Compressed row (CRS) or compressed column storage (CCS) formats\(^{79}\) can be used to store the sparse \( A \) matrix very efficiently.

Solutions of an overdetermined system of linear equations have no meaning in an absolute sense; therefore, at least one of the energy levels needs to be fixed. SNs are rooted graphs; therefore, it is an obvious choice to fix the value of the lowest energy level, the root (or one of the roots of the PCs), and fix it to zero. Due to the nature of the experimental SNs, they may contain energy levels that have no path to the root; therefore, these energy levels must be identified before setting up the matrix equation. The fastest way to identify the components of the SN and to select nodes belonging to the same component approaches similar to MARVEL, including those developed by Flaud et al.\(^{75}\) and Tashkun et al.\(^{32}\)
of the SN is the DFS algorithm, which significantly outperforms the Dijkstra algorithm for this task.

Both iterative and direct linear solver algorithms can be used to determine the MARVEL energy levels, i.e., X. A considerable advantage of the direct methods is that the elements of the inverse matrix can be determined analytically. With \( A^{-1} \), the uncertainties (one standard deviation) of the energy levels can be computed as \( \epsilon_i \approx \sqrt{A_{ii}^{-1}} \). A considerable disadvantage of the direct methods is that they are much slower than the iterative algorithms. Because \( A \) is a symmetric positive definite matrix, we can use a sparse-adaptive iterative algorithm. Because the direct methods is that they are much slower than the special type of the Cholesky decomposition. The robust weighting algorithm is especially well suited for the iterative adjustment of the uncertainties of the measured transitions. If approximate uncertainties are sufficient, like during the MARVEL iterations up to the final one to improve the experimental uncertainties, it is possible to use the preconditioned conjugate gradient method.

4.1. Magic Numbers. As mentioned above, the absolute values of the MARVEL energy levels can only be determined if at least one energy level is fixed in the given SN. Most of the SNs, as prescribed by quantum mechanical selection rules, have more than one PC. One of these PCs contains the lowest energy level as root, whose value is chosen to be zero, with zero uncertainty. MARVEL cannot determine the absolute values of the lowest energy level of the other PCs, so we must link the PCs by so-called "magic numbers" to have absolute values for all the energies.

The value of the magic number can be estimated on the basis of empirical and/or theoretical considerations. The magic number can usually be deduced from a highly accurate empirical effective Hamiltonian. The experimental SN can also be used for the determination of magic numbers. This facility is provided by the observation that many molecules have degenerate energy pairs where the members of the pairs belong to different PCs. This may happen at only relatively high excitations and energies, but many of these excited states are usually amenable to experiments. These near degeneracies can be ascertained from accurate variational nuclear motion computations. Adding these artificial, zero-frequency "transitions" to the experimental SN the PCs become artificially connected and a refinement process can be initiated yielding the magic number. According to our experience, the empirical magic number satisfies the experimental accuracy what is required during a MARVEL analysis.

4.2. Calibration. The databases that allow the execution of a MARVEL analysis contain all the measured transitions of a given molecule available from the literature, which means that these transitions have been measured over several decades under widely different experimental conditions, including pressure and temperature differences, using different spectrometers and different calibration standards. When these data are combined into a single database, systematic differences can be identified if several groups reported precise values with different accuracy and uncertainty estimates for the same transitions, yielding a multiedge SN. Inconsistencies may occur due to mistakes of different origin, but some of the inconsistencies, especially in the case of Fourier transform spectroscopy (FTS) measurements, are due to the use of different calibration standards recommended at different times.

The incorrectly calibrated FTS transitions can easily be corrected by applying a single multiplicative recalibration factor. However, this factor needs to be determined. An example is provided by the high-resolution spectroscopic data measured by Guelachvili for \( \text{H}_2\text{O} \), and a couple of its isotopologues, and reported in 1983 in the 1066–2296 cm\(^{-1}\) region. These data were revised about a decade later, when Guelachvili et al. introduced a calibration factor of 0.99999977, improving substantially the accuracy of the measured lines. Although the deviation of this factor from 1.0 appears to be small, many FTS measurements have a relative accuracy considerably better than 10\(^{-7}\), under ideal conditions this can be 10\(^{-9}\)–10\(^{-10}\).

The base of the MARVEL determination of this multiplicative calibration factor is the minimization of the root-mean-square deviation between the FTS transitions scaled with a given factor and the MARVEL predicted transitions. Using the MARVEL calibration protocol we could determine basically the same calibration factor for \( \text{H}_2\text{O} \) as determined by Guelachvili et al. A similar situation was observed during the MARVEL studies of several other molecules.

4.3. Conflict of Highly Accurate Lines Measured for \( \text{H}_2\text{O} \). An interesting feasible application of network theory and MARVEL concerns the "planning" of experiments, i.e., identification of unmeasured transition(s), which can result in new energy levels or solve an existing conflict among different measurements. A case in mind is when experimentalists claim higher accuracy than apparently their measurement has (or the accuracy is lowered substantially due to special circumstances, like the presence of strongly overlapping lines).

The following example considers a conflict between two sets of highly accurate measured transitions of the molecular ion \( \text{H}_2\text{O}^+ \). In 2013, two different groups studied the \( v_2 \) band of \( \text{H}_2\text{O}^+ \) and published highly accurate lines; however, their measured frequencies did not agree within the published uncertainties. Table 5 contains selected transitions from these two sources that exemplify the problem. The notation applied for the transitions of Table 5 is explained neither here nor in the table, the original sources should be consulted for this purpose. MARVEL and the SN approach in itself cannot solve this apparent conflict of the two measurements, i.e., cannot select the "more accurate" transitions, because the energy levels involved in the conflict are not members of cycles. Nevertheless, the SN approach can predict, using the appropriate selection rules, accurate transitions that could create cycles among the energy levels involved in the problems. If it was feasible to measure these predicted transitions with an accuracy similar to those of the measurements of refs. and 109, then the cycle(s) created by the new measurements could fix the energy levels resolving the conflict. Figure 10 shows the
problem and the solution schematically. The blue and the red lines are the measured transitions (see also Table 5), and the green lines show the predicted MARVEL lines that appear in the region of feasible measurements and can create a cycle that could prove the accuracy of either one or the other original experimental study. During the revision of this paper a study from Asvany et al.110 came to our attention, reporting improved, highly accurate lines for H3⁺. The results of ref 110 seem to indicate that McCall et al.108 significantly overestimated the accuracy of their measurements, the results of ref 110 came to our attention, reporting uncertainties of the energy levels may reflect this weakly connected property. This figure also nicely shows the large number of branches almost always characteristic of experimental SNs.

5. SUMMARY AND CONCLUSIONS

Driven by the need of scientific and engineering applications, information systems containing line-by-line high-resolution spectroscopic data have become ubiquitous during the last 20 years or so. These information systems contain rovibronic transitions for a considerable number of small molecules, selected on the basis of the need of the applications, usually containing two to six atoms. The largest of the computed line lists have up to 10^10 transitions,111,112 clearly calling for “big data” techniques to generate, store, validate, distribute, and utilize the information. It must be emphasized that although quantum chemical computations are able to yield the complete set of the required line-by-line information even for semirigid pentatomic molecules, like 12CH4,112 the accuracy of the line positions is limited and thus should be replaced by much more accurate experimental data whenever they are available. As emphasized in this article, to make maximum use of the experimental spectroscopic line list information, the grand list of measured spectroscopic transitions of a molecule should be viewed as a weighted, undirected, and rooted graph, a spectroscopic network (SN). The vertices of the SN are the energy levels (these are principally independent of the type of measurement providing them), whereas the edges are the spectroscopically allowed transitions and the weights are the transition intensities (both depend on the type of measurement

Figure 10. Part of the measured spectroscopic network of H3⁺ showing pictorially a conflict due to the accuracy of two sets of measurements: the red transitions were measured by 13HoPeJeSi108 and the blue ones by 13WuLiLiLi.109 The green, unmeasured transitions within the same measurable wavenumber range should help to solve the conflict of the two sets of measurements once and for all.

Figure 11. Clauset–Newman–Moore (CNM) clusters of the experimental spectroscopic network of the 12C2 molecule.
performed). Obviously, transition intensities have a crucial role in determining the structure of SNs and different spectroscopic techniques yield SNs with drastically different topologies.

Experiments yield relatively small multiedge random graphs: the largest experimental SN studied, that of ortho- and para-$\text{H}_2^{16}\text{O}$, contains about 20000 energy levels and 200000 transitions, of which about 100000 are unique. A given first-principles computation of rovibronic energy levels and spectra results in a very large deterministic simple graph.

The network-theoretical view of the results of high-resolution spectroscopy experiments yielding rovibronic transitions and energy levels, after introducing the concept of spectroscopic networks (SN) and once one understands the structure of the SNs, offers several concepts and tools toward the complete characterization of the related rovibronic energies and spectra, some of which can be summarized as follows:

1. SNs of simple molecules usually contain more than one principal component (PC), as required by nuclear spin statistics. PCs are giant components of the spectroscopic networks. Experimental spectroscopic networks occasionally contain floating components, whose energy levels are not attached to any of the energy levels of the PCs. Unification of the components of an experimental SN is important to obtain “absolute” energy values for the vertices of components disjoint from the PCs.

2. The degree distribution of all experimental SNs investigated turns out to be free of a scale. The scale-free property of the overall network degree distribution of SNs thus established leads to the useful concept of hubs, i.e., the emergence of a relatively few energy levels with a relatively large number of transitions. Note that this statement is independent of the assumed degree distribution of SNs; it only relies on the heavy-tail distribution observed in all cases. The established existence of hubs provides design ideas for highly useful spectroscopic measurements. For example, accurate measurement of transitions involving the least well characterized hubs of the experimental SN leads straightforwardly to a more accurate list of levels and lines.

3. The fact that the PCs of experimental SNs are giant components can be explained by the observation that all experimental SNs studied exhibit heavy tails in their degree distribution with the exponent of the assumed fitted power-law distribution of about 2.2. This scaling index means, according to modeling studies, that the SNs are allowed to have giant components and small ones occur only occasionally. This is a useful property of spectroscopic networks as it ensures that most of the rovibronic energy levels can be obtained from a diverse set of experimental measurements.

4. All the experimental spectroscopic networks investigated turn out to be bipartite. This is another important property of SNs. It reflects the fact that the parity of the energy levels has to change during experimentally measurable one-photon transitions. The bipartite nature of SNs means that SNs contain only even-membered cycles, the smallest possible cycle involves four energy levels. Bipartiteness allows, for example, for a simple partial check of the correctness of the labels of the lines listed in spectroscopic databases.

5. Detailed investigation of first-principles SNs show that, although the great majority of the transitions is extremely weak, it is possible to find a few relatively strong transitions for almost all energy levels. This suggests that with relatively standard spectroscopic techniques almost all of the bound rovibronic energy levels can be determined through measurement of transitions they are involved in.

6. Detailed comparison of measured and computed hubs helps to determine the weakest links within an experimental SN, i.e., those transitions that are least well determined and whose accurate knowledge is most important to ensure the overall accuracy of the lines and levels involved in the SN.

7. It seems that the PageRank order of the hubs can be significantly different from their degree order. PageRank provides the more useful measure as within the PageRank ordering system hubs are preferentially connected with hubs.

8. The scale-free spectroscopic networks are robust against random removal of nodes. The robust structure of the experimental spectroscopic networks investigated means that they have a small diameter, resulting in the ultrasmall-world property of SNs.

9. The matrix representations of SNs, involving the adjacency, and the combinatorial and normalized Laplacians, can be used to learn a number of details about the structure of the experimental spectroscopic networks. Most of the times the same information can be obtained by other means, but for the clustering of SNs the normalized Laplacian seems to offer the best opportunity. It is tempting to believe that a rovibronic energy level is experimentally well determined if it is part of a cycle. Clustering methods, some relying on matrix representations of SNs, help to detect small communities of rovibronic energy levels in which the uncertainties of the energy levels should reflect this weakly connected property.

10. Besides model Hamiltonians with a relatively small number of parameters, another opportunity to reduce spectroscopic data to manageable size is the conversion of rovibronic transitions into rovibronic energy levels. Minimum-weight spanning trees, where the weights are defined by transition intensities, comprise the minimum amount of transition information needed to represent the experimentally available energy levels.

11. In a high-resolution study a spectrum of a polyatomic molecule is converted into a list of labeled energy levels. Among other techniques, this can be achieved very efficiently via the use of the minimum-weight spanning tree, as it identifies the most intense and thus practically most useful spectral features.

Probably the most important application of the concept of spectroscopic networks is their use within the MARVEL (Measured Active Rotational–Vibrational Energy Levels) procedure yielding rovibronic energy levels, referenced to a selected zero level, from measured transitions involving them. The MARVEL code is not only completely general and can be applied to any molecule but also is very fast, allowing on-the-fly analysis of arbitrary experimental SNs and experimental spectra. MARVEL has been used to study the experimental energy level structure of 15 molecules and yielded tens of...
thousands of highly accurate energy levels. In favorable cases MARVEL allows the “experimental” determination of the energy difference between the roots of the principal components of experimental SNs, through degeneracies of (highly excited) energy levels belonging to different principal components. MARVEL also facilitates the calibration of Fourier-transform spectroscopy studies, an important feature when transitions from many different experimental sources must be used together. As shown for example for H$_3^+$, the network-theoretical view of molecular spectra helps to understand conflicts of existing experiments and to propose new experiments to resolve the contradictions.

Spectroscopic networks, perhaps as part of active databases, are expected to become an intrinsic part of the description of high-resolution spectra of molecules. Nevertheless, investigation of large-scale SNs, containing hundreds of thousands of nodes and hundreds of millions of links calls for further high-resolution spectra of molecules. Nevertheless, investigations, like the scale-free and properties, hubs, network dynamics, self-organization, robustness, and attack/error tolerance, are all relevant when experimental (and to some extent first principles) spectroscopic networks are characterized.

**AUTHOR INFORMATION**

Corresponding Author
*A. G. Császár* Phone: +36-1-372-2929. E-mail: csaszar@chem.elte.hu.

Notes
The authors do not compete financial interest.

**ACKNOWLEDGMENTS**

The project was supported by NKFIH (grant no. NK83583). The authors gratefully acknowledge the many useful discussions related to the topic of this article with Dr. Marianna Bolla, Dr. Beáta Faller, Prof. László Lovász, Dr. Edit Mátys, Dr. Georg Mellau, Prof. Jonathan Tennyson, and Dr. János Tóth. The support received via the COST Action CM1405, MOLIM: Molecules in Motion, is also acknowledged.

**REFERENCES**

(1) Euler, L. The seven bridges of Königberg. *Commentarii academiae scientiarum Petropolitanae* 1741, 8, 128–140.


(3) Interestingly, graph theory owes its name to chemistry, as the "graphical notation" or "chemicograph" of chemical species abundantly used in the 19th century, and pioneered by Brown and Kekulé, was abbreviated by the noted mathematician John Joseph Sylvester to "graphs" in 1878, see: Sylvester, J. J. Chemistry and algebra.


(44) Mellau, G. Ch Complete experimental rovibrational eigenergies of HNC up to 3743 cm−1 above the ground state. J. Chem. Phys. 2010, 133, 164303.
(45) Mellau, G. Ch Complete experimental rovibrational eigenergies of HCN up to 6880 cm−1 above the ground state. J. Chem. Phys. 2011, 134, 234303.
Article

The Journal of Physical Chemistry A


(101) It must be emphasized that if we want to examine certain properties of the SN, such as the diameter, the Dijkstra algorithm appears to be an excellent choice.


