J Math Chem (2012) 50:2342–2350
DOI 10.1007/s10910-012-0033-7

ORIGINAL PAPER

Determining polyhedral arrangements of atoms using PageRank

Matthew Hudelson · Barbara Logan Mooney · Aurora E. Clark

Received: 30 January 2012 / Accepted: 23 May 2012 / Published online: 28 June 2012

Abstract Polyhedral representations of the geometric arrangements of atoms and molecules is a pervasive tool in chemistry for understanding chemical bonding and electrostatic interactions. Yet the structural organization within very large systems is often difficult to quantify. In this work, we illustrate that PageRank, when combined with the chemical constraints of a system, can be used to uniquely identify the polyhedral arrangements of atoms and molecules. The PageRank algorithm can be used on any network that can be represented as a graph: a mathematical object where individual points, or vertices, are joined by edges. It is thus well-suited for chemical systems where atoms (considered vertices) are connected to each other via chemical bonding (considered edges) or other forces. This has been implemented in a recently reported series of R-scripts, *moleculaRnetworks*, and the example provided herein illustrates that the polyhedral arrangement of solvent molecules about a solute results in a unique *PR* value for the solute and enables rapid identification of the local geometry in the condensed medium. More generally *PR* can be used as a chemoinformatic tool to search for specific structural patterns within any database of geometric configurations.

Keywords PageRank · Graph theory · Solvation · Polyhedra

M. Hudelson (✉)
Department of Mathematics, Washington State University, Pullman, WA 99164, USA
e-mail: mhudelson@wsu.edu

B. L. Mooney
Department of Chemistry and Biochemistry, University of Arizona, Tucson, AZ 85721, USA

A. E. Clark (✉)
Department of Chemistry, Washington State University, Pullman, WA 99164, USA
e-mail: auclark@wsu.edu

1 Introduction

Statistical mechanical (SM) simulations provide a wealth of data regarding the ensemble of atomic positions that are populated for a system under a set of experimental conditions. Most often chemists are interested in the local geometric arrangements of atoms and molecules, whose tendency to form polyhedral structures is a well known result of the underlying physics associated with chemical bonding and electrostatic interactions. The interpretation of SM data for this purpose is often done in a direct (visual) manner and few algorithms exist for an automated approach toward structural characterization. In a recent work, we reported the *moleculaRnetworks* series of R-scripts for the post-processing of SM data and the study of solvent structure about solutes [1]. Applications of this code have investigated the solvent shell structure and exchange processes of mono-, di-, and trivalent metal cations in water [2]. An essential aspect of *moleculaRnetworks* is the utilization of the PageRank (*PR*) algorithm to identify the connectivity and organization of atomic/molecular networks. *PR* is best known for its implementation in the Google internet search engine to assign numerical weighting to each element of a hyperlinked set of documents [3]. However, the PageRank algorithm can be used on any network that can be represented as a graph: a mathematical object where individual points, or vertices, are joined by edges. The success of *moleculaRnetworks* is in the application of *PR* to graphs formed from "connecting the dots" between atomic positions as a means of identifying the polyhedral arrangements of those atoms. In this work, we demonstrate the uniqueness of the *PR* to a given graph. In combination with the known behavior of chemical systems and the natural constraints therein, the graphs formed in *moleculaRnetworks* become restricted to those representing convex polyhedra, and thus the *PR* becomes a new identifying tool for characterizing local geometry.

2 Method

As reported in [3], the normalized PageRank formula that is implemented for ranking internet web-pages is:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L\left(p_j\right)}$$

where the set of pages is $\{p_1, p_2, \ldots, p_N\}$, the value $PR(p_i)$ denotes the page rank of page p_i, $L\left(p_j\right)$ is the number of outgoing links from page p_j, and $M(p_i)$ is the set of pages that link to page p_i. If a page p_j has no outgoing links (is a "sink" page), then we assume $L\left(p_j\right) = N$ and $p_j \in M(p_i)$ for all i. This emulates a surfer picking a random page when they are finished with a page that contains no links. The value of d represents the probability that a surfer follows a link on the page he/she is on. The term $(1-d)/N$ represents the probability of a surfer beginning a new surfing session on page p_i; the $1-d$ factor is the probability of beginning a new surfing session and the $1/N$ factor represents the probability of choosing page at random. To adopt this convention to a system of atoms, the index $\{p_1, p_2, \ldots, p_N\}$ merely becomes the list of atoms $\{1, 2, \ldots, N\}$, the page rank of atom i becomes $PR(i)$ and and the number of connections from the atoms j that are connected to i is $L(j)$. Thus, the *PR* of atom i is determined not only by the number of connection it has to atoms j, but also each j's connectivity to other atoms in the network/graph. The definition of a "connection" is modulated by the radial cutoff from atom i that obeys the definition of the chemical phenomena under investigation. If only the immediate geometry of the atom i is being considered, the cutoff would include only atoms that are directly bonded to i. Alternatively, non-bonded interactions may be considered, such as H-bonding, in which case a typical H-bond distance could be used. In the case of statistical mechanical data, the cutoff values can be determined from the pair distribution function (PDF) between the atom types of interest.

Let $\vec{1}$ represent the all-ones vector and M represent the matrix whose entries are

$$M_{i,j} = \begin{cases} 1/L(j), & \text{if } i \in M(j) \\ 0, & \text{otherwise.} \end{cases}$$

Then the vector $\vec{r} = [PR(1)\,, PR(2)\,, \ldots, PR(n)]^T$ is a solution to $\vec{r} = \frac{1-d}{N}\vec{1} + dM\vec{r}$. Alternatively, $\vec{r}$ solves $(I - dM)\,\vec{r} = \frac{1-d}{N}\vec{1}$. Since M is a stochastic matrix, the moduli of its eigenvalues are bounded above by 1 and so is invertible for any $0 \leq d < 1$. Thus, $\vec{r} = \frac{1-d}{N}\,(I - dM)^{-1}\,\vec{1}$. The above discussion allows us to conclude that the PageRank vector $\vec{r}$ represents a unique solution to the equation $\vec{r} = \frac{1-d}{N}\vec{1} + dM\vec{r}$. Thus, if we are able to demonstrate that a given vector solves the equation, then we will not have to be concerned that some other vector is a solution as well. This amounts to the demonstration that the *PR* of atom i, whose connected atoms form a polyhedral geometry about i is distinct for different n-vertex polygons. While this will be shown not to be the case for simple point charges, the known physical constraints of chemical systems cause this to be true in practice (*vide infra*).

In chemical systems the linking is symmetric, i.e., if i is connected to j, then j is connected to i. Let's also assume there are no isolated (non-bonded) atoms. In this case, we let G be a graph whose vertices are the atoms i and where two vertices are joined by an edge if and only if the atoms they represent are "connected". Suppose A is the adjacency matrix for G and $\tilde{D}$ is the diagonal matrix whose diagonal entries are $\tilde{D}_{ii} = 1/d_i$, the reciprocal of the degree d_i of vertex i. Then it is a routine computation to verify that $M = A\tilde{D}$. Another routine computation verifies that if $\vec{q} = [d_1, d_2, \ldots, d_N]^T$ then $M\vec{q} = A\tilde{D}\vec{q} = A\vec{1} = \vec{q}$. Thus, the vector containing the degrees of the vertices is an eigenvector of M with eigenvalue 1. Ultimately, this means that the solution vector $\vec{r} = \vec{r}(d)$ to the equation $\vec{r} = \frac{1-d}{N}\vec{1} + dM\vec{r}$ must satisfy

$$\vec{r}(1) = \frac{1}{d_1 + d_2 + \cdots + d_N} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix}.$$

We obtain the following result:

Theorem 1 *Suppose atomic connections are symmetric and suppose $\vec{r}(d)$ is the solution vector to both equations $\vec{r} = \frac{1-d}{N}\vec{1} + dM\vec{r}$ and $\vec{r} = \frac{1-d}{N}\vec{1} + dK\vec{r}$. Then the underlying graphs for M and K must have the same degree sequences.*

It is not safe to conclude that the underlying graphs are isomorphic. Consider any k- regular graph G with n vertices. We form the graph G' by adjoining to G another vertex u_{n+1} and forming edges joining u_{n+1} to every vertex in G'. The matrix for G' will have entries

$$M_{r,c} = \begin{cases} 0, & \text{vertices } u_r \text{ and } u_c \text{ are not adjacent} \\ 1/(k+1), & \text{vertices } u_r \text{ and } u_c \text{ are adjacent, } c < n+1 \\ 1/n, & c = n+1, r < n+1 \end{cases}$$

In block form,

$$M = \begin{bmatrix} (k+1)^{-1}A & n^{-1}\vec{1} \\ (k+1)^{-1}\vec{1}^T & 0 \end{bmatrix}$$

where A is the adjacency matrix for G. We observe that $\vec{1} \in \mathbb{R}^n$ is an eigenvector for with eigenvalue k.

Now, we consider determining the value of α that solves the equation

$$dM\begin{bmatrix} \alpha\vec{1} \\ 1-n\alpha \end{bmatrix} + \frac{1-d}{n+1}\begin{bmatrix} \vec{1} \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha\vec{1} \\ 1-n\alpha \end{bmatrix}.$$

Expanding the left-hand side, we obtain

$$\begin{aligned} dM\begin{bmatrix} \alpha\vec{1} \\ 1-n\alpha \end{bmatrix} + \frac{1-d}{n+1}\begin{bmatrix} \vec{1} \\ 1 \end{bmatrix} &= d\begin{bmatrix} (k+1)^{-1}A & n^{-1}\vec{1} \\ (k+1)^{-1}\vec{1}^T & 0 \end{bmatrix}\begin{bmatrix} \alpha\vec{1} \\ 1-n\alpha \end{bmatrix} + \frac{1-d}{n+1}\begin{bmatrix} \vec{1} \\ 1 \end{bmatrix} \\ &= d\begin{bmatrix} \alpha\,(k+1)^{-1}A\vec{1} + (1-n\alpha)\,n^{-1}\vec{1} \\ \alpha n\,(k+1)^{-1} \end{bmatrix} + \frac{1-d}{n+1}\begin{bmatrix} \vec{1} \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \left(d\alpha k\,(k+1)^{-1} + d\,(1-n\alpha)\,n^{-1} + (1-d)\,(n+1)^{-1}\right)\vec{1} \\ d\alpha n\,(k+1)^{-1} + (1-d)\,(n+1)^{-1} \end{bmatrix} \end{aligned}$$

which means we seek α such that

$$\begin{aligned} d\alpha k(k+1)^{-1} + d(1-n\alpha)\,n^{-1} + (1-d)\,(n+1)^{-1} &= \alpha \\ d\alpha n(k+1)^{-1} + (1-d)\,(n+1)^{-1} &= 1-n\alpha. \end{aligned}$$

Solving the first equation for α yields

$$\alpha\left(1 + d - \frac{dk}{k+1}\right) = \frac{d}{n} + \frac{1-d}{n+1}$$

$$\alpha \frac{1+d+k}{k+1} = \frac{n+d}{n(n+1)}$$
$$\alpha = \frac{(n+d)\,(k+1)}{n\,(n+1)\,(d+k+1)}.$$

As a check, solving the second equation for α yields

$$\alpha \left(\frac{dn+n(k+1)}{k+1} \right) + \frac{1-d}{n+1} = 1$$
$$\alpha \frac{n(d+k+1)}{k+1} = \frac{n+d}{n+1}$$
$$\alpha = \frac{(n+d)\,(k+1)}{n\,(n+1)\,(d+k+1)}$$

which verifies our answer from the first equation. Thus, with $\vec{r} = \begin{bmatrix} \alpha\vec{1} \\ 1-n\alpha \end{bmatrix}$ where $\alpha = \frac{(n+d)(k+1)}{n(n+1)(d+k+1)}$, the equation

$$dM\vec{r} + \frac{1-d}{n+1}\vec{1} = \vec{r}$$

is solved. From the previous discussion, this is the unique solution to this equation. We note that the entries of $\vec{r}$ sum to 1, and so we have produced a PageRank vector. We observe that this PageRank vector is the same if we begin with any regular graph on n vertices, so two such graphs cannot be distinguished by this process. As an example, suppose G_1 is a six-membered ring with atom i in the center and G_2 is has two disjoint three-membered rings above and below the plane of i as in Fig. 1.
Letting A_1 and A_2 be their adjacency matrices and M_1 and M_2 be the associated M-matrices, we have

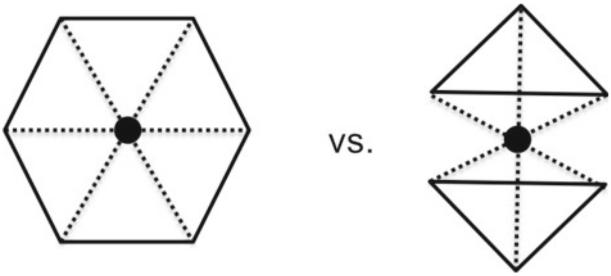


Fig. 1 Cyclic six-membered ring versus two disjoint triangles connected to a single point/atom in a regular graph

$$M_1 = \begin{bmatrix} 0 & 1/3 & 0 & 0 & 0 & 1/3 & 1/6 \\ 1/3 & 0 & 1/3 & 0 & 0 & 0 & 1/6 \\ 0 & 1/3 & 0 & 1/3 & 0 & 0 & 1/6 \\ 0 & 0 & 1/3 & 0 & 1/3 & 0 & 1/6 \\ 0 & 0 & 0 & 1/3 & 0 & 1/3 & 1/6 \\ 1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/6 \\ 1/3 & 1/3 & 1/3 & 1/3 & 1/3 & 1/3 & 0 \end{bmatrix} \quad \text{and}$$

$$M_2 = \begin{bmatrix} 0 & 1/3 & 1/3 & 0 & 0 & 0 & 1/6 \\ 1/3 & 0 & 1/3 & 0 & 0 & 0 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 0 & 0 & 1/6 \\ 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/6 \\ 0 & 0 & 0 & 1/3 & 0 & 1/3 & 1/6 \\ 0 & 0 & 0 & 1/3 & 1/3 & 0 & 1/6 \\ 1/3 & 1/3 & 1/3 & 1/3 & 1/3 & 1/3 & 0 \end{bmatrix}.$$

We have $n = 6, k = 2$ and $\alpha = \frac{(n+d)(k+1)}{n(n+1)(d+k+1)} = \frac{d+6}{14d+42}$. The vector

$$\vec{r} = \frac{1}{14d+42} \begin{bmatrix} d+6 \\ d+6 \\ d+6 \\ d+6 \\ d+6 \\ d+6 \\ 8d+6 \end{bmatrix}$$

solves both equations $dM_1\vec{r} + \frac{1-d}{7}\vec{1} = \vec{r}$ and $dM_2\vec{r} + \frac{1-d}{7}\vec{1} = \vec{r}$.

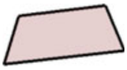
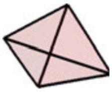
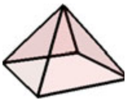
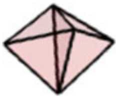
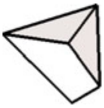
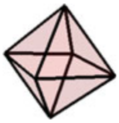
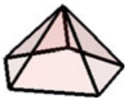
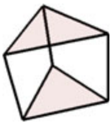
The above example is of course only valid for structures that have exactly the same degree sequence. Examples where this may occur include: (a) identical isolated molecules or clusters where two configurations are energetically favorable (as proposed in Fig. 1) or, (b) a perfectly repeating periodic crystal of both symmetries. In both of these cases, the definition of a "connection" is well-defined, the radial cutoff from the central atom terminates beyond the bounds of the cluster in the first case, and is dictated by the periodic boundary of the crystal in the second. However, in most chemical applications such equivalent degree sequences are incredibly difficult to achieve. When more realistic chemical systems come under investigation, where the vertices of the polygon are themselves part of an extended network of connectivity, and where an ensemble of configurations are being considered (as is the case for statistical mechanical data) the probability of having the same degree sequence for different polyhedral arrangements becomes quite small. The chemical constraints of the system also help to ensure that the *PR* of an atom becomes a unique identifier of local structure in practical applications.

As an example, let us consider a prototypical system for analysis using the *moleculaRnetworks* scripts which incorporate *PR* as a means to identify molecular structures from statistical mechanical simlations. In the case of an ion immersed in water, H_2O molecules form concentric shells of solvation about the ion and the ion-dipole interaction in the immediate vicinity of the ion (the first solvation shell) causes the H_2O to be highly organized. The classical approach for characterizing the solvation shell is by means of a pair distribution function (PDF) between the ion center and the O-atom of water, which will have at least one well-defined peak that when integrated, produces the "coordination number" (CN) of the ion. The first peak in the PDF indicates that the waters in the first solvation shell are distributed about a maximally probable radial distance. Moreover, since the solvent molecules repel one another equally they will be distributed roughly uniformly on a sphere of radius r , and thus may be mapped to a convex polyhedron. Planar solvation shells as might be inferred from the hexagon in Fig. 1 are nonsensical in a chemical system such as this (e.g. they would have a huge cavitation energy).

We are interested in the *PR* of the ion and as such consider the connection between the first solvation shell O-atoms and the ion, as well as the intermolecular connections between O-atoms. Thus, the relevant graphs for this chemical system are star graphs. Each graph in *moleculaRnetworks* begins as a star graph, with connections between each solvent molecule (as defined by atomic position) and the central solute. Upper and lower bounds to the length of this connection are established by examination of the pair distribution function. The lower bound is dictated by the repulsive part of the potential that describes the ion-water interaction (governed by the hard-sphere radius of each atom), while the upper bound is the boundary between the first and second concentric solvation shells about the ion (which appear as separate peaks in the pair distribution function). Next, edges are formed between water O-atom vertices if the distance between them is less than or equal to a cutoff distance, which we have taken to be the side of the largest cube fitting into a sphere of radius r with r less than the first minimum of the pair distribution, but greater than the first maximum. This does two things. First, it forms the polyhedral skeleton without edge crossings through the center of the graph or behind other water vertices, as such edges would not be present in a convex polyhedron. Second, it limits the number of edges had by an individual vertex, such that we need not worry about excessively large numbers of edges. The use of the cutoff distance as a chemical constraint for edge formation can also alter the number of edges for each vertex. For example, if a planar hexagon solvation shell were to be present in solution and at the same time a solvation shell with three H_2O arranged above and below the plane of the ion, as in Fig. 1, then the cutoff distance is chosen to be large enough that at least one edge is formed between the top and bottom triangles, as illustrated in Table 1. Thus, as implemented in *moleculaRnetworks*, the disjoint triangles in Fig. 1 become identified as a trigonal biprismatic polygon when the top and bottom triangles are sufficiently close to one another, and the two solvation environments will yield distinct *PR* values for the central ion due to different degree sequences now being present. As another example, consider the trigonal prism and octahedron (Table 1) which both have CN = 6 for the central ion. Provided that the number of edges at each vertex is different for the two forms—for the prism, 3, and for the antiprism, 4—the introduction of the ion "breaks" the regularity of the graphs, altering their degree sequences and *PR*, thus allowing the two forms to be distinguished.

The calculated *PR* of the central ion based upon these chemical constraints and the limitations of a star graph are shown in Table 1 for four- to six-vertex polyhedra. Note

Table 1 PageRank of the central ion for N-vertex polyhedra (N = 4–6)

Number of vertices	Polygon name	Shape	PageRank
4	Square		0.2441558
4	Tetrahedron		0.200000
5	Square pyramid		0.1892430
5	Trigonal bipyramid		0.1772388
5	Wedge		0.2035064
6	Octagon		0.1636142
6	Pentagonal pyramid		0.1822820
6	Trigonal prism		0.1929308

that in instances where atomic positions fluctuate such that their distance is greater than the cutoff, the polyhedra database within *moleculaRnetworks* also includes the *PR*s of graphs with one or two edges missing, so that structures can be matched when a water is slightly beyond the cutoff. To list them all is beyond the scope of this work, but we note one example in the “wedge,” which is not a proper convex polyhedron, but rather an arrangement gleaned from the MD data of multiple ions in water solvent, in which one water vertex has only two edges. As an example, in our prior work [1,2], we performed an analysis of 1ns of MD data of Na$^+$ion immersed in 216 TIP3P water molecules. Through the application of *moleculaRnetworks*, the geometry in the first solvation shell, Na$(H_2O)_5^+$, was identified in 99.4 % of configurations. The 5-coordinated ion was found in the square pyramidal geometry 53.6 % of the time, and in the triangular bipyramid and “wedge” 9.3 and 36.6 % of the time. This indicates that nearly all of the 5-vertex polyhedra are identifiable.

3 Conclusions

Chemical structures or networks that have exactly the same degree sequence are shown to have the same PageRank. However, exploitation of chemical constraints can ensure the construction of non-regular graphs with unique *PR*. For the example provided, matching the *PR* of the graphs from statistical mechanical data to the *PR* in a database of ideal polyhedra (as is done in *moleculaRnetworks*), can reduce the complicated problem of identification of solvent organization to the easy application of a graph-theoretic algorithm. The example provided herein presents a useful tool for the analysis of SM data of a wide array of systems, from biological solutes like proteins, to reactive species in solution within chemistry and chemical engineering. Keep in mind that the *PR* is unique for different degree sequences of a system, and thus can be used as a fingerprint of solvent organization even if the solvent is does not form convex polyhedra as in the example presented. Indeed, more generally *PR* can be used as a chemoinformatic tool to search for specific structural patterns within any database of geometric configurations. In this respect, the *PR* and the connectivity information contained within the M matrix could be used in a similar way to the Morgan algorithm that was so instrumental in the early development of the Chemical Abstracts Service for assigning unique identifying labels to different chemical structures [4].

Acknowledgments This work was supported in part by a grant from the Department of Energy, Basic Energy Sciences, Heavy Element program (DE-SC0001815), as well as the National Science Foundation, award # 0848346.

References

1. B.L. Mooney, L.R. Corrales, A.E. Clark, J. Comp. Chem. (2012). doi:10.1002/jcc.2917 (early view)
2. B.L. Mooney, L.R. Corrales, A.E. Clark, J. Phys. Chem. B **116**, 3387 (2012)
3. S. Brin, L. Page, in *Proceedings of the 7th International Conference on the World Wide Web (WWW).* eds. by Enslow, P. H., Ellis, A. (Elsevier, Amsterdam, 1998), p. 107
4. H.L. Morgan, J. Chem. Doc. **5**, 107 (1965)