Chemical graph theory

Jacob Kautzky
MacMillan Group Meeting
April 3, 2018
Chemical graph theory

What is graph theory?

Historical uses of chemical graph theory

- Isomer counting
- Chemical bonding
- Kinetics

Modern uses of chemical graph theory
A graph is defined as a non empty set of vertices and a set of edges

Order of the graph \( (N) = \) # of vertices
Size of the graph \( (M) = \) # of edges
Degree of a vertex \( (D) = \) # of edges incident with a vertex

\[
\sum_{i=1}^{N} D(i) = 2M
\]

\[N - M + F = 2\]

Walks, trails, paths, and cycles

**Walk** - an alternate sequence of vertices and edges, beginning and ending with a vertex

- **Open** - starts and ends at different vertices
- **Closed** - starts and ends at the same vertex

**Length** ($l$) - number of occurrence of edges in a walk

**Trail** - a walk where all edges are distinct

**Path** - a walk where all vertices are distinct

**Distance** ($d$) - the length of the shortest path between 2 vertices

**Cycle** - a path that starts and ends at the same vertex

**Girth** - length of the smallest cycle in a graph
Connected vs disconnected and simple, general, and multigraphs

**Connected graph** - $\exists$ a path between any 2 vertices

**Disconnected graph**

**Simple graph** - a graph without loops or multiple edges between two vertices

**Multigraph** - a graph that allows multiple edges between 2 vertices

**General graph** - a graph that allows multiple edges and loops

Forests and trees

**Forest** - an acyclic graph

**Tree** - a connected acyclic graph

**Spanning tree** - a graph subgraph that includes all vertices of a graph with the minimal number of edges to remain connected

Types of graphs — regular graphs

A regular graph is a graph where every vertex has the same valency

- A graph where each vertex has $D = N - 1$
- Denoted by $K_N$
- $M = \binom{N}{2}$

- A graph where each vertex has $D = 2$
- Denoted by $C_N$
- A cycle is termed even/odd if $N$ is even/odd

$M = 0.5 \times N \times D$

Vertex coloring

Assigning colors to vertices so that no two adjacent vertices have the same color

Chromatic number ($k$) - the minimal number of colors required to color a graph

$k = 3$

$k = 5$

$k = 4$

Bipartite graph - a graph where the vertex set can be split into $V_1$ and $V_2$ where every edge connects a vertex in $V_1$ to one in $V_2$

- A graph is a bipartite graph if and only if every cycle is of even number
- All bipartite graphs are 2 colorable

Complete bipartite graph $K_{s,u}$ - every vertex in $V_1$ is joined to every vertex in $V_2$

Planar graphs

A graph is planar if it can be drawn such that no edges intersect

Two graphs are isomorphic if there exists a 1:1 mapping from one onto the other (i.e. they are identical, but drawn differently)

Identifying isomorphic graphs without sampling all $N!$ mappings is challenging and remains a problem in graph theory

Planar graphs

A graph is planar if it can be drawn such that no edges intersect

Planar

A graph is planar if it does not contain a subgraph homeomorphic to $K_5$ or $K_{3,3}$

Not planar

Two graphs are homeomorphic if they can be obtained from the same graph by inserting new vertices of valency 2 into its edges

All planar graphs are 4-colorable

**Euler Circuit – the Königsberg bridge problem**

Is it possible to cross every bridge exactly once and start and end at the same spot?

**Graphical Model**

**Examples of Euler Circuits**

**Examples of Euler Paths**

- An Euler circuit is a path that starts and ends at the same vertices and traverses every edge exactly once.
  - Only occurs if every vertex is of even degree.

- An Euler path is a path that traverses every edge exactly once, but doesn’t start and end at the same vertex.
  - Only occurs if every vertex besides two are of even degree.

**Hamiltonian circuit – the icosian game**

Is it possible to start at one vertex of a dodecahedron, travel to every other vertex on the polyhedra, and then return to the initial vertex without visiting any vertex besides the starting one twice?

**Graphical Model**

**Hamiltonian Circuit**

- A circuit that visits every vertex exactly once
- A mathematical formula for if a circuit exists has yet to be developed
- If the $D(i) \geq N/2$ for every vertex $i$, then a circuit exists

---

Digraphs, line graphs, and weighted graphs

**Digraph**

**Weighted Graph**

**Line graph**

Graphs can be converted to matrices and polynomials

Adjacency matrix

Distance matrix

Characteristic Polynomial

\[ P(G) = \det (\lambda I - A) \]

Applying graph theory to chemistry

**Molecular Structures**
- Vertices represent atoms
- Edges represent bonds
- Weighted edges can represent double bonds or C-X bonds

**Polymers**
- Vertices represent building blocks
- Edges represent connections

**Kinetics**
- Vertices represent intermediates
- Edges represent pathways
- Frequently digraphs

**And many more...**
- Aromaticity
- Depicting orbitals and electrons
- NMR analysis
- Crystals and clusters
- Mapping reaction space

Chemical graph theory

What is graph theory?

Historical uses of chemical graph theory

- Isomer counting
- Chemical bonding
- Kinetics

Modern uses of chemical graph theory
How can we go about counting the number of isomers possible for a given carbon number in a systematic fashion?

- C = 1
- C = 2
- C = 3
- C = 4
- C = 5
- C = 6
- C = 7
Isomer counting – the Caylee approach

Recursive approach

Counted the number of centric and bicentric trees

Developed methods for counting rooted and unrooted trees and then limited them to 4 vertices

Successfully counted the number of C1-C11 alkanes

Rather tedious process

Gave incorrect answers for C12 and C13 alkanes

Isomer counting – the Henze-Blair approach

Counting the number of alcohols on acyclic alkanes

Let $T_N$ be the number of alcohols of carbon $N$ and $p_N$, $s_N$, and $t_N$ be the number of primary, secondary and tertiary alcohols respectively.

Then $T_N = p_N + s_N + t_N$

$p_N = T_{N-1}$

$s_N = \begin{cases} 
T_1 * T_{N-2} + T_2 * T_{N-3} + \ldots + T_{(N-2)/2} * T_{N/2} & \text{if } N = \text{even} \\
T_1 * T_{N-2} + T_2 * T_{N-3} + \ldots + T_{(N-3)/2} * T_{(N+1)/2} + (1/2) * T_{(N-1)/2} * [1 + T_{(N-1)/2}] & \text{if } N = \text{odd}
\end{cases}$

Isomer counting – the Henze-Blair approach

A tertiary alcohol can be imagined to arise from combining 3 alkyl radicals $R_i$, $R_j$, and $R_k$ with $i$, $j$, and $k$ carbons respectively to the COH group.

$$t_N = \begin{cases} 
\sum T_i \cdot T_j \cdot T_k & i > j > k; \ i + j + k = N - 1 \\
+ \frac{1}{2} \sum T_i \cdot (1 + T_i) \cdot T_k & i = j; \ i + j + k = N - 1 \\
+ \frac{1}{6} T_i \cdot (T_i + 1) \cdot (T_i + 2) & i = j = k; \ i + j + k = N - 1 
\end{cases}$$

This approach has been expanded to deal with structural saturated hydrocarbons, unsaturated hydrocarbons, alkynes and even stereoisomeric alcohols

Isomer counting – the Pólya enumeration approach

Takes into account symmetry operations

Produces a polynomial that allows for isomer enumeration

\( f \) and \( g \) are equivalent if and only if there exists a permutation \( \alpha \) such that \( \alpha(f) = g \)

\[ Z(A) = \frac{1}{|A|} \sum_{\alpha \in A} \prod s_r^{j_{\alpha}} \]

How many ways are there to substitute a benzene ring with an R group \( X \) times?

Point group: \( D_6 \)

<table>
<thead>
<tr>
<th>Symmetry Operation</th>
<th>Permutation</th>
<th>Cycle index term</th>
</tr>
</thead>
<tbody>
<tr>
<td>One identity ( E )</td>
<td>(1) (2) (3) (4) (5) (6)</td>
<td>( s_1^6 )</td>
</tr>
<tr>
<td>Two 6-fold rotations +/- ( C_6 )</td>
<td>(123456), (165432)</td>
<td>( 2s_6^1 )</td>
</tr>
<tr>
<td>Two 3-fold rotations +/- ( C_3 )</td>
<td>(153)(264), (135)(246)</td>
<td>( 2s_3^2 )</td>
</tr>
<tr>
<td>One vertical 2-fold rotation ( C_2 )</td>
<td>(14)(25)(36)</td>
<td>( s_2^3 )</td>
</tr>
<tr>
<td>Three in plane binary axes ( C_2 ) that bisect 0 carbons</td>
<td>(16)(25)(34), (12)(36)(45), (14)(23)(56)</td>
<td>( 3s_2^3 )</td>
</tr>
<tr>
<td>Three in plane binary axes ( C_2 ) that bisect 2 carbons</td>
<td>(1)(4)(26)(35), (3)(6)(15)(24), (2)(5)(13)(46)</td>
<td>( 3s_1^2 s_2^2 )</td>
</tr>
</tbody>
</table>

Isomer counting – the Pólya enumeration approach

Takes into account symmetry operations

Produces a polynomial that allows for isomer enumeration

\( f \) and \( g \) are equivalent if and only if there exists a permutation \( \alpha \) such that \( \alpha(f) = g \)

Zyklenzeiger: 
\[
Z(A) = \frac{1}{|A|} \sum_{\alpha \in A} \prod_{r} s_{r}^{j(\alpha)}
\]

Point group: \( D_{6} \)

How many ways are there to substitute a benzene ring with an R group \( X \) times?

\[
Z(D_{6}) = \frac{1}{12} \{ s_{1}^{6} + 3s_{1}^{2}s_{2}^{2} + 4s_{3}^{3} + 2s_{3}^{2} + 2s_{6}^{1} \}
\]

substituting

\[
s_{y}^{\frac{z}{y}} = (1 + x^{y})^{2}
\]

\[
Z(D_{6}) = \frac{1}{12} \left[ (1 + x)^{6} + 3 \cdot (1 + x)^{2} \cdot (1 + x^{2})^{2} + 4 \cdot (1 + x^{2})^{3} + 2 \cdot (1 + x^{3})^{2} + 2 \cdot (1 + x^{6}) \right]
\]

There are 3 unique isomers with 4 R substituents

\[
Z(D_{6}) = 1 + x + 3x^{2} + 3x^{3} + 3x^{4} + x^{5} + x^{6}
\]

Isomer counting – the Pólya enumeration approach

Takes into account symmetry operations

Produces a polynomial that allows for isomer enumeration

\[ f \text{ and } g \text{ are equivalent if and only if there } \exists \text{ a permutation } \alpha \text{ such that } \alpha(f) = g \]

\[ \text{Zyklenteiger: } Z(A) = \frac{1}{|A|} \sum_{\alpha \in A} \prod_{r} s_{r}^{j(\alpha)} \]

How many ways are there to substitute a benzene ring with an R group X times?

Point group: D_6

\[ Z(D_6) = 1 + x + 3x^2 + 3x^3 + 3x^4 + x^5 + x^6 \]

Isomer counting – the Pólya enumeration approach

Takes into account symmetry operations

Produces a polynomial that allows for isomer enumeration

\[ f \text{ and } g \text{ are equivalent if and only if there } \exists \text{ a } \alpha \text{ such that } \alpha(f) = g \]

Zykeningzer: \[ Z(A) = \frac{1}{|A|} \sum_{\alpha \in A} \prod s_{j}^{k(\alpha)} \]

How many ways are there to substitute a benzene ring with an R group X times?

Point group: D\textsubscript{6}

This method has been extended to isotopic isomers, cyclic molecules, benzenoid hydrocarbons, porphyrins, chiral and achiral alkenes, ferrocenes, clusters, and inorganic structures, among others

Isomer counting – the N-tuple code

Assigns each tree a unique code
Isomorphic compounds have the same code
Have been used for generation and enumeration of acyclic graphs

Start from most substituted vertex
Remove vertex and incident edges, then examine the subtrees
Produce the lexicographically largest code

Chemical graph theory

What is graph theory?

Historical uses of chemical graph theory

- Isomer counting
- Chemical bonding
- Kinetics

Modern uses of chemical graph theory
NMR spectroscopy can probe both structure and dynamics of biomolecules at atomic resolutions.

- Methyl-TROSY can be utilized to study protein complexes up to 1 MDa in molecular weight.
- A major challenge to this approach is the need to match resonances in the NMR with specific atoms.
- Generally obtained by either monitoring assignments from smaller intact proteins or individually changing residues and observing the NMR perturbations.

Would it be possible to use graph theory to simplify this process?

Methyl Assignment by Graph Matching

NMR assignment of methyl peaks in large molecules via MAGMA

NMR assignment of methyl peaks in large molecules via MAGMA

Kinetic Analysis with Prim’s algorithm

Minimum spanning tree – a subset of edges that connect a graph together with the minimal possible weight

Prim’s algorithm finds minimum spanning by repeatedly adding the cheapest vertex to the spanning tree

Kinetic Analysis with Prim’s algorithm

\[
\text{cat. } \left[ \text{Rh}(I)((S)\text{-binap})L_2 \right]^+ + \text{Me}_2\text{HS}\text{Me} \rightarrow \text{Me}_2\text{H}_{\text{S}}\text{H}_{\text{Me}}\text{Me} \quad \text{THF, 60 °C, 15 h, 97% ee}
\]

Kinetic analysis with Prim’s algorithm

- Bolded edges are part of the spanning tree
- Green paths represent earlier in the pathway and red later

Kinetic analysis with Prim’s algorithm

Dissociative

- Bolded edges are part of the spanning tree

- Green paths represent earlier in the pathway and red later

Can we use graph theory to predict reactions?

High-throughput computer-based reaction predictions (HTRP) can be used for de novo drug design, virtual chemical space exploration, and to predict if retrosynthesis disconnections are feasible.

Can we use this to predict reactions?

- Constructed a knowledge graph from 14.4 million reactants and 8.2 million binary reactions.
- Vertices represent reactants and reaction conditions with edges connecting reactants to reaction conditions and other reactants.
- Predict reactions, products and reaction conditions.
- Correctly predicted the right product with 67.5% accuracy on 180,000 reactions that were withheld from the data set.
- Able to predict the products and reaction conditions of ‘novel’ reactions with decent efficiency.

Can we use graph theory to predict reactions?

High-throughput computer-based reaction predictions (HTRP) can be used for de novo drug design, virtual chemical space exploration, and to predict if retrosynthesis disconnections are feasible.

Can we use this to predict reactions?

- Constructed a knowledge graph from 14.4 million reactants and 8.2 million binary reactions.
- Vertices represent reactants and reaction conditions with edges connecting reactants to reaction conditions and other reactants.
- Predict reactions, products and reaction conditions.
- Correctly predicted the right product with 67.5% accuracy on 180,000 reactions that were withheld from the data set.
- Able to predict the products and reaction conditions of ‘novel’ reactions with decent efficiency.
Can we use graph theory to predict reactions?

Can we use graph theory to predict reactions?

Would it be possible to develop a reaction between these 2 components?

Can we use graph theory to predict reactions?

**Predicted**

\[
\text{Ir(ppy)}_2(\text{dtbbpy})\text{PF}_6 \\
\text{Pd(Ph}_3\text{)}_4, \text{AcOH, MeCN}
\]

**Discovered**

\[
\text{Ir(ppy)}_2(\text{dtbbpy})\text{PF}_6 \\
\text{Pd(Ph}_3\text{)}_4, \text{HCO}_2\text{H, MeCN}
\]


Applying graph theory to chemistry

**Molecular Structures**
- Vertices represent atoms
- Edges represent bonds
- Weighted edges can represent double bonds or C-X bonds

**Polymers**
- Vertices represent building blocks
- Edges represent connections

**Kinetics**
- Vertices represent intermediates
- Edges represent pathways
- Frequently digraphs

**And many more...**
- Aromaticity
- Depicting orbitals and electrons
- NMR analysis
- Crystals and clusters
- Mapping reaction space
Questions

Figure 13. A plot showing the annual number of papers published in the area of chemical graph theory for the years 1970-1986. Note the annual growth rate of around 25% for this period.