Relevant Cycles in Chemical Reaction Networks

Petra M. Gleiss

Institut für Theoretische Chemie und Molekulare Strukturbiologie,
Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
`pmg@tbi.univie.ac.at`

Peter F. Stadler[†]

Institut für Theoretische Chemie und Molekulare Strukturbiologie,
Universität Wien
Santa Fe Institute
1399 Hyde Park Road, Santa Fe, NM 87501, USA
`studla@tbi.univie.ac.at`, `stadler@santafe.edu`

Andreas Wagner

Dept. of Biology, University of New Mexico,
167A Castetter Hall, Albuquerque, NM-817131-1091,
Santa Fe Institute
`wagnera@unm.edu`

David A. Fell

School of Biological & Molecular Sciences
Oxford Brookes University, Oxford OX3 OBP, U.K.
`daf@brookes.ac.uk`

We characterize the distributions of short cycles in a large metabolic network previously shown to have small world characteristics and a power law degree distribution. Compared with three classes of random networks, including Erdős-Rényi random graphs and synthetic small world networks of the same connectivity, both the metabolic network and models for the chemical reaction networks of planetary atmospheres have a particularly large number of triangles and a deficit in large cycles. Short cycles reduce the length of detours when a connection is clipped, so we propose that long cycles in metabolism may have been selected against in order to shorten transition times and reduce the likelihood of oscillations in response to external perturbations.
KEYWORDS: cycle basis, small world network, metabolic network, planetary atmosphere

[†] Address for correspondence

Glossary

BA	Albert-Barabasi scale free network
CFC	chloro-fluoro-carbon compounds
`Ecoli1`	The model of the E. coli core metabolism from (Wagner and Fell, 2000)
ER	Erdős-Rényi random graph
HC	Hydrocarbon reaction network
MCB	Minimal cycle basis
SWk	Watts-Strogath Small World network with connections to the k nearest neighbors

1. Introduction

The rapidly increasing amount of molecular data on genes, proteins and metabolic pathways enhances our capability to study cell behaviour. The understanding of the molecular logic of cells requires the ability to analyze metabolic processes and gene networks in qualitative and quantitative terms. In this contribution we shall focus on global features of metabolic networks that are likely to have evolved and thus set the networks of cellular chemistry apart from abiotic reaction networks.

Recent surveys, in particular (Jeong *et al.*, 2000; Wagner and Fell, 2000; Fell and Wagner, 2000), have revealed that metabolic reaction networks belong to the class of small world networks in the wider sense: they have a diameter that is much smaller than what one would expect for an uncorrelated random graph with the same number of vertices and edges.

Small world networks have received considerable attention since the seminal paper by Watts and Strogatz (Watts and Strogatz, 1998). In a recent paper (Amaral *et al.*, 2000), Amaral *et al.* present evidence that there are (at least) three structurally different classes of networks that are distinguished by the distribution $P(d)$ of the vertex degrees d:

(a) *Single Scale Networks* with a sharp distribution of vertex degrees exhibiting exponential or Gaussian tails. This class includes also the Erdős-Rényi model of uncorrelated random graphs (Erdős and Rényi, 1960; Bollobás, 1985).

(b) *Scale Free Networks* with a power law distribution $P(d) \sim d^{-\gamma}$. A simple model for this type of networks was introduced recently by Barabási *et al.* (Barabási and Albert, 1999; Barabási *et al.*, 1999). Metabolic networks (Wagner and Fell, 2000; Jeong *et al.*, 2000) and food-webs (Montoya and Solé, 2000) belong to this class.

(c) *Broad Scale Networks* for which $P(d)$ has a power-law regime followed by a sharp cut-off, e.g. exponential or Gaussian decay of the tail. An example is the movie-actor network described in (Watts, 1999)

The most common model of graph evolution, introduced by Erdős and Rényi (Erdős and Rényi, 1960), assumes a fixed number $n = |V|$ of vertices and assigns edges independently with a certain probability p (Bollobás, 1985). In many cases ER random graphs turn out the be quite different from a network of interest. The Watts-Strogatz (Watts and Strogatz, 1998) model of small world networks

starts with a deterministic graph, usually a circular arrangement of vertices in which each vertex is connected to k nearest neighbors on each side. Then edges are "rewired" (in the original version) or added (Newman and Watts, 1999; Newman *et al.*, 2000) with probability p. We shall consider the latter model for $k = 1$, denoted SW1 below, which corresponds to adding random edges to a Hamiltonian cycle. Both ER and SW1 graphs exhibit an approximately Gaussian degree distribution.

The other extreme is scale-free BA model (Barabási and Albert, 1999; Barabási *et al.*, 1999) with a degree distribution of the form $P(d) \sim d^{-3}$: Starting from a small core graph, at each time step a vertex is added together with m edges that are connected to each previously present vertex k with probability

$$\Pi(k) = d(k) \big/ \sum_j d(j) \,, \qquad (1.1)$$

where $d(j)$ is the degree of vertex j. A recent extension of the model allows the tuning of the scaling exponent γ in the range $2 \leq \gamma \leq 3$ (Albert and Barabasi, 2000).

Much of the literature discusses small world networks in terms of the average path length between two vertices (Newman *et al.*, 2000) or of the network's clustering coefficient (Herzel, 1998; Barrat and Weigt, 2000) which measures how close the neighborhood of a each vertex comes on average to being a complete subgraph (clique) (Watts and Strogatz, 1998). In this contribution we consider the small cycle of small world networks in detail. This approach is motivated by the following two observations:

Recent work on the spread of epidemics on a small world network (Pandit and Amritkar, 1999) emphasizes the importance of "far-reaching" edges. The idea is that clipping a far edge will force a (relatively) long detour in the network. Hence it is these edges that are responsible for the small diameter of the graph G. We shall see in section 2 that detours are intimately related to the cycles in the graph. In particular, we describe the connection between cycles in directed and undirected models and argue that the collection of *relevant cycles* is the appropriate mathematical object for our purposes. In section 3 we briefly outline the relationship between the cycle structure of a reaction network and *Chemical Flux Analysis*. In the following section the distribution of triangles and longer relevant cycles is discussed for uncorrelated random graphs as well as for small world models. In section 6 we compare two classes of chemical reaction networks here: (1) Metabolic networks in which all reactions are mediated by specific enzymes, and (2) the reaction networks of planetary atmospheres which lack specific catalysis. A discussion of our results and open problems concludes this contribution.

2. Detours, Cycles, and Circuits

Let us look at detours in graphs in more systematic way. Throughout this paper we will represent a network as a simple (unweighted, undirected) graph $G(V, E)$ with vertex set V and edge set E. A *cycle* in G is a closed path which meets each of its vertices and edges exactly once. The length of a cycle C, i.e., the number of its vertices or edges, is denoted by $|C|$. With each edge $e \in E$ we
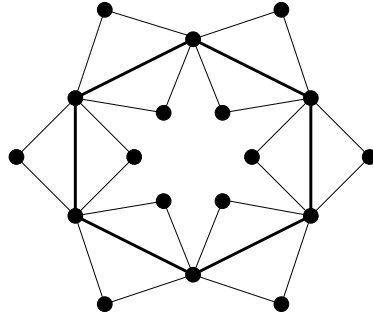
Figure 1: $\mathcal{S}(G)$ consists of the twelve triangle only. The hexagon (bold edges), however, is crucial for the network structure since the local information conveyed by the triangles does not allow the reconstruction of the hexagonal overall-structure.

can associate the set $\mathcal{S}(e)$ containing the shortest cycles in G that go through e. It is easily verified that a far edge in the sense of (Pandit and Amritkar, 1999) is an edge that is not contained in a triangle. In other words, e is a far edge if and only if $\mathcal{S}(e)$ does not contain a triangle. The cycles $C \in \mathcal{S}(e)$ determine the shortest detours (which have length $|C| - 1$) when e is removed from the graph.

It seems natural to consider the set $\mathcal{S}(G) = \bigcup_{e \in E} \mathcal{S}(e)$ of shortest cycles of all edges in G and to study e.g. their length distribution. However, as the l.h.s. example in Fig.1 shows, the shortest cycles $\mathcal{S}(G)$ do not convey the complete information about the graph. Additional cycles appear to be relevant, such as the hexagon in Figure 1. It cannot be reconstructed from the collection of triangles, but it determines the diameter (maximal distance between two vertices) of the graph.

A formal treatment of cycles in undirected graph conveniently starts with the notion of the *cycle space*. The set of all subsets of E forms an $|E|$-dimensional vector space over $\{0, 1\}$ (with addition and multiplication modulo 2). Vector addition in this *edge space* is given by symmetric difference $X \oplus Y = (X \cup Y) \setminus (X \cap Y)$. The *cycle space* $\mathcal{C}$ consisting of all cycles and edge-disjoint unions of cycles in G is a particularly important subspace of the edge space (Chen, 1971). The dimension of the cycle space is the *cyclomatic number* $\nu(G) = |E| - |V| + c(G)$, where $c(G)$ is the number of connected components of G.

The notion of a cycle space is readily extended to directed graphs $\vec{G}(V, A)$ with vertex set V and *arc set A*. A *circuit* is a directed closed path on the vertex set of $\vec{G}$ such that each arc and each vertex (with the exception of the starting point) is visited exactly once. With each circuit one associates a vector $C \in \mathbb{R}^{|A|}$ with entries $C_a = +1$ if the a is in the circuit and its orientation agrees with the orientation of the path, $C_a = -1$ if the orientation of arc and circuit are opposite and $C_a = 0$ if the arc a is not part of the circuit, see Fig.2 for an example. The *circuit space* is the vector space over $\mathbb{R}$ generated by (the vector representations of) all circuits on $\vec{G}$. Note that if C is a circuit then $-C$ is the circuit with the opposite orientation.

The length of a cycle or a circuit is the number of its edges or arcs, respectively. In terms of the vector representation we have $|C| = \sum_a |C_a|$. For each collection
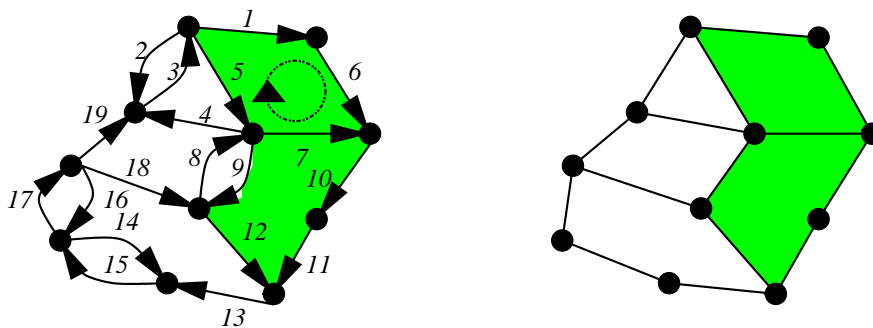
Figure 2: The circuit C delimiting the shaded region in the directed graph $\vec{G}$ on the l.h.s. and the orientation indicated by the arrow has the vector representation

$$(+1, 0, 0, 0, -1, +1, 0, 0, -1, +1, +1, -1, 0, 0, 0, 0, 0, 0, 0).$$

The r.h.s. shows the graph G°, obtained from $\vec{G}$ by one edge of each 2-cycle and ignoring the direction of the arcs. We have for example $R = \{2, 8, 15, 16\}$ or $R' = \{3, 8, 14, 17\}$ as "reverse arcs" that are omitted in passing from $\vec{G}$ to G°.

$\mathcal{B}$ of cycles or circuits we define the length

$$\ell(\mathcal{B}) = \sum_{C \in \mathcal{B}} |C|. \tag{2.1}$$

This definition is meaningful for both cycles of undirected graphs and circuits of directed graphs. A *minimum cycle basis* (MCB) is a cycle basis with minimum length. MCBs have the property that their longest cycle is at most as long as the longest cycle of any basis of $\mathcal{C}$ (Chickering *et al.*, 1994). A MCB therefore contains the salient information about the cycle structure of a graph in its most compressed form. It appears natural to consider the cycle structure of a graph in terms of its MCBs, in particular, because the following theorem establishes a strong connection between the shortest detours discussed in the previous section and minimal cycle bases.

THEOREM 2.1. (Stepanec, 1964; Zykov, 1969)
Each cycle $C \in \mathcal{S}(G)$ of an undirected graph is contained in a minimal cycle basis.

In general, graphs do not have unique MCBs. In fact, the known classes of graphs with unique MCB have a very simple structure: they are outer-planar (Leydold and Stadler, 1998), Halin graphs (Stadler, 2000), or certain series-parallel graphs (McKee, 2000). However, the distribution of cycle sizes is the same in all MCBs of a graph G. More precisely:

THEOREM 2.2. (Stepanec, 1964; Chickering *et al.*, 1994)
Suppose $\mathcal{M}$ is a MCB of G containing n_k cycles of length k. Then every MCB of G has exactly n_k cycles of length k.

A cycle is called *essential* if it is contained in all minimal cycle bases. In particular, a cycle is essential if it is the unique shortest cycle through some edge
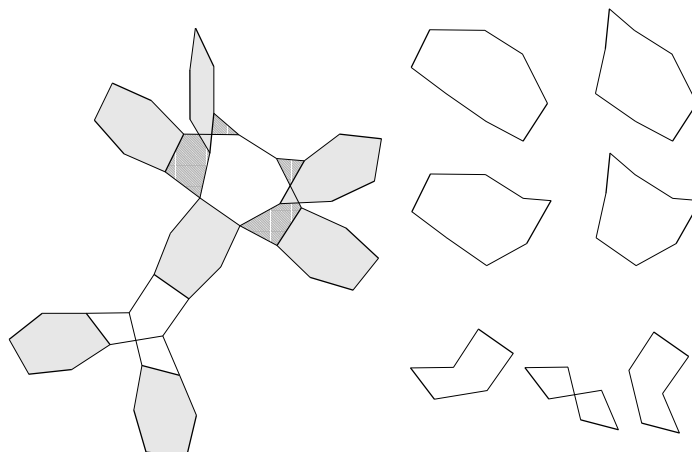
Figure 3: Organic carbon compounds may exhibit elaborate polycyclic structures. The example shown here is Compound 8 from (Kammermeier *et al.*, 1996) (aromatic "double bonds" are indicated by thick lines). The nine essential cycles are marked with gray shades. There are two groups of relevant non-essential cycles: four 8-rings and three 6-rings. A minimal cycle basis contains two of the three 6-rings and one of the four 8-rings.

(Gleiss and Stadler, 1999). This condition is not necessary, however. An example of a graph with a non-unique MCB indicating the essential cycles is given in Fig. 3.

The main difference between undirected and directed graphs arises from the fact that in the directed case we can have two arcs connecting the same pair of vertices, one in each direction. Such pairs of arcs form circuits of length 2 which we shall call 2-cycles. With a directed graph $\vec{G}(V, A)$ we may associate in a natural way an undirected graph $G^\circ(V, E)$ with the same vertex set V and an edge set E obtained by ignoring the direction of the arcs. In the case of cycles of length two we identify one of the arc with the undirected edge and omit the other. For a given directed graph $\vec{G}$ the underlying undirected graph G° is uniquely determined. The sets $E \subseteq A$ of retained edges and the corresponding set $R = A \setminus E$ of omitted "reverse edges", on the other hand, are not unique. It depends obviously on the (arbitrary) choice of one of the two arcs of each 2-cycle. We shall sometimes write G°_R to emphasize the fact that the correspondence of A and E depends on R.

It is clear that C and $-C$ are always linearly dependent, hence a basis of the circuit space cannot contain both orientations of the same circuit. On the other hand, since the change of orientation does not change the length of the circuit, we can always replace C by $-C$ in a basis. We can therefore disregard the orientation of the circuit and consider it only as the set of its arcs.

This observation allows us to obtain a simple correspondence between the circuit space of a directed graph $\vec{G}$ and the cycle space of is underlying undirected graph G°:

THEOREM 2.3. *A collection $\mathcal{Q}$ of circuits in $\vec{G}$ is a minimal length basis of the*

circuit space if and only if $\mathcal{Q} = \mathcal{M} \cup \mathcal{D}$, where $\mathcal{M}$ is a minimal cycle basis of G_R°
for some R, and $\mathcal{D}$ is the set of 2-cycles in $\vec{G}$.

PROOF. The circuit space of $\vec{G}$ has the dimension $\nu(\vec{G}) = |A| - |V| + \mathrm{c}(G^\circ)$
where $\mathrm{c}(G^\circ)$ denotes the number connected components of G°, i.e., the number
of weak components of $\vec{G}$, see e.g. (Berge, 1985, Thm.3.4). It is clear that any
basis $\mathcal{B}$ of the cycle space of G° forms a linearly independent set in the circuit
space of $\vec{G}$. Furthermore, the set $\mathcal{B} \cup \mathcal{D}$ is linearly independent since each 2-cycle
in $\mathcal{D}$ contains an arc that is neither part of another 2-cycle nor part of a cycle
in $\mathcal{B}$. We have
$|\mathcal{B} \cup \mathcal{D}| = \nu(G^\circ) + |\mathcal{D}| = |E| - |V| + \mathrm{c}(G^\circ) + |R| = |A| - |V| + \mathrm{c}(G^\circ) = \nu(\vec{G})$,
hence $\mathcal{B} \cup \mathcal{D}$ is a basis of the circuit space.
Now we can use the well-known fact that the circuits form a matroid (see e.g.
(Horton, 1987; Hartvigsen and Mardon, 1993)) to obtain minimal length cycle
bases by means of the greedy algorithm from the set $\mathcal{Z}$ of all circuits. The crucial
observation is that $\mathcal{Z}$ consists of the double edges, i.e. 2-cycles, and of circuits
that are obtained from the cycles of G by omitting a particular combination of
reverse edges R. We have already seen that a basis may contain a circuit only in
one of the two orientations. The shortest circuits in $\mathcal{Z}$ are the $|\mathcal{D}|$ 2-cycles, which
are linearly independent. The greedy criterion thus implies that each of them is
contained in all minimal length bases. All other circuits are strictly longer and
by construction belong to cycles in G°. Let C and C' be two circuits belonging
to the same cycle C° with different sets omitted reverse edges R and R'. Then C'
can be written as a linear combination of C and the 2-cycles formed from $R \cup R'$.
Thus $\mathcal{D} \cup \{\mathcal{C}, \mathcal{C}'\}$ is linearly dependent and hence at most one circuit belonging
to a cycle C can be part of the basis. Conversely, if the greedy algorithm selects
a circuit C then it can be replaced by any circuit C' with the same underlying
cycle C° and any R without affecting length or linear independence.
Thus the greedy algorithm effectively selects a minimum cycle basis of G°. We
have already argued in the first part of the proof that $\mathcal{D}$ is extended to a circuit
basis by any cycle basis of the undirected graph G°. Thus we obtain a minimal
circuit bases by extending $\mathcal{D}$ with a minimal cycle basis $\mathcal{M}$ of G°. $\square$

Theorem 2.3 shows that as far as the cycle structure is concerned we may safely
neglect the direction of the arcs. As an immediate consequence Theorems 2.1 and
2.2 are valid for minimum length bases of the circuit space as well. As far as the
cycle structure of a network is concerned, we are thus justified in restricting
ourselves to the underlying undirected graph G°.

The main shortcoming of MCBs for the characterization of graphs is the fact
that the MCB is not unique in general. A natural way to avoid ambiguities is to
consider the union of all minimum cycles bases, also known as the set $\mathcal{R}(G)$ of
relevant cycles. The term "relevant" is justified by two important properties of
$\mathcal{R}(G)$: (i) a cycle is relevant if and only if it cannot be written as an $\oplus$-sum of
shorter cycles (Vismara, 1997), and (ii) the shortest cycles through an edge are
relevant, i.e., $\mathcal{S}(G) \subseteq \mathcal{R}(G)$ as an immediate consequence of Theorem 2.1. Con-
sequently, the composition of $\mathcal{R}(G)$ in terms of number and length distribution
of cycles is an important characteristic of a graph. The numerical studies below

make use of Vismara's (Vismara, 1997) algorithm for computing $\mathcal{R}(G)$, which is based on Horton's MCB algorithm (Horton, 1987).

3. Flux Analysis in Chemical Networks

Let us now return to chemical networks. Because it is germane to their functional analysis, we first point out a nexus between graph representations of metabolic network, and metabolic flux analysis (MFA), the most generic framework to analyze the biological function of metabolic networks.

The key ingredient of MFA is the *stoichiometric matrix* $\mathbf{S}$. Its entries are the stoichiometric coefficients s_{kr}, i.e., the number of molecules of species k produced ($s_{kr} > 0$) or consumed ($s_{kr} < 0$) in each reaction r. Reversible reactions are entered as two separate reactions in most references. In general, additional "pseudo-reactions" are added to describe the interface of the metabolic reaction network with its environment.

The dynamics of the concentration of metabolite k may be generally described by

$$\frac{\mathrm{d}c_k}{\mathrm{d}t} = \sum_r s_{kr} J_r - \mu(t)c_k \tag{3.1}$$

where the flux J_r depends on the kinetic properties of the participating enzymes, on the concentrations of metabolites and on environmental parameters such as temperature and pH. The enzymes are generally subject to complex regulations by inhibition and activation. The assumption of a steady state and neglecting the dilution as a consequence of low concentrations of intermediates yields the homogeneous, time-independent system of linear equations

$$\mathbf{S}J = \vec{o} \tag{3.2}$$

for the flux vector J. Consequently, the steady state flux vectors are elements of the null-space $\mathsf{Null}(\mathbf{S})$. Using the constraint that we have must have $J_r \geq 0$ for each reaction r, we see that J is a steady state flux vector iff $J \in \mathsf{Null}(\mathbf{S}) \cap \mathbb{R}_+^{|V|}$. The extremal rays of this cone are usually called the *elementary flux modes* and are closely associated with the relevant metabolic pathways, see e.g. (Clarke, 1988; Heinrich and Schuster, 1996; Fell, 1997; Schilling *et al.*, 2000; Edwards and Palsson, 2000; Schuster *et al.*, 2000) for further details on MFA.

It is not hard to see that if all reactions are mono-molecular, then $\mathbf{S}$ is the incidence matrix of a directed graph: $s_{kr} = 1$ for the single product k formed in reaction r and $s_{kr} = -1$ for the single metabolite used in reaction r, i.e., $\mathbf{S}$ is the incidence matrix of the digraph $\vec{G}$ whose vertices are the chemical species and whose edges denote the reactions. Such networks were studied already in the 1960s (Balaban *et al.*, 1966). It is well known that x is element of the cycle space of $\vec{G}$ if and only if $\mathbf{S}x = \vec{o}$, i.e., the circuit space of $\vec{G}$ is $\mathsf{Null}(\mathbf{S})$ (Bollobás, 1998). The stationary flux vectors are therefore cycles of $\vec{G}$.

In general, $\mathbf{S}$ represents a *directed hypergraph* (Zeigarnik, 2000). Equivalently, one may use a bipartite graph in which one class of vertices represents the substrates and the other class of vertices denotes the reactions. Arcs point from the educts to the reaction node and from the reaction node to the products, Fig. 4.

Figure 4: Representations of the reaction $NO_2 + O_3 \rightarrow NO_3 + O_2$ in hypergraph form drawn as the equivalent directed bipartite graph (l.h.s) and as part of a substrate graph (r.h.s).

A very simple graph representation of chemical networks, which is sufficient for our purposes, is the *substrate graph* Σ introduced in (Wagner and Fell, 2000). Its vertices are the molecular compounds (substrates); two substrates k and l are adjacent in Σ if they participate in the same reaction r. The substrate graph is a straight-forward approximation of the directed hypergraph representing **S**: a directed hyper-edge is replaced by a clique on the same set of vertices. As a consequence, the stationary flux vectors are closely related to the cycles of substrate graph.

We consider undirected substrate graphs because directed graphs would not properly represent the propagation of perturbations: even for irreversible reactions the product concentration may affect the reaction rate, for instance by product occupancy of the enzyme's active site; this in turn affects the substrate concentration. Thus, perturbations may travel backwards even from irreversible reactions. A similar argument for considering undirected graphs can be derived from metabolic control theory (Sen, 1991). A number of more complicated graph representations are discussed e.g. in the book (Temkin *et al.*, 1996).

Note that **S** does not identify the input and output metabolites. This information is added in the form of additional "I/O-vertices" and "pseudo-reactions" representing flux in and out of the reaction network in MFA applications, see e.g. (Fell, 1997). The corresponding extension of the cycle space of the network graph is the vector space spanned by all cycles of the reaction network and all paths connecting pairs of "I/O-vertices" (Hartvigsen, 1993). The generalization of the notion of relevant cycles to this extended vector space will be explored in detail elsewhere.

4. Triangles

It is clear that all triangles in a graph are relevant, since a triangle is necessarily a shortest cycle through each of its edges. Hence $|\mathcal{R}(G)| \geq \Delta$, where Δ denotes the number of triangles in G. We expect $\langle\Delta\rangle_{\mathrm{ER}} = \binom{n}{3}p^3$ triangles in an ER random graph with edge-drawing probability p. For the SW1 graphs we obtain a similar expression:

$$\langle\Delta\rangle_{\mathrm{SW1}} = np + n(n-4)p^2 + \frac{1}{6}n(n^2 - 9n + 20)p^3 . \qquad (4.1)$$

The MCB will therefore consist almost exclusively of triangles if $\Delta \gg \nu(G)$. The average vertex degree is $d = 2|E|/n = p(n-1)$ for ER and $d = 2 + p(n-3)$ for
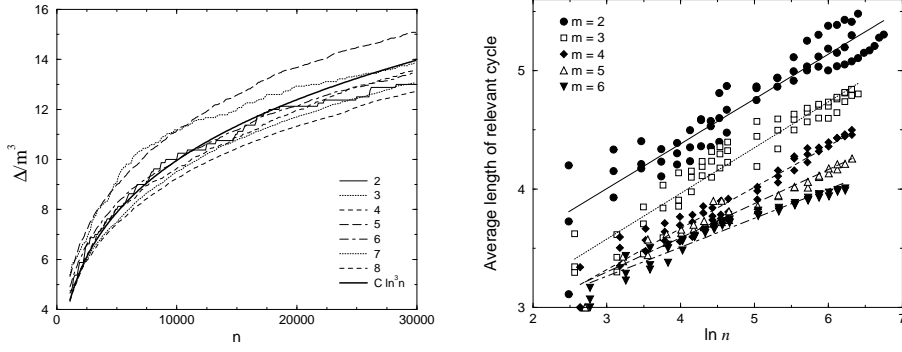
Figure 5: Cycles in the BA model.
L.h.s.: triangles in BA models with different values of m.
R.h.s.: mean length of a relevant cycle in BA networks.

SW1, resp. Assuming that n is large we expect to find only triangles in $\mathcal{R}(G)$ for $d \gg \sqrt{3n}$. Numerical simulations show that this is indeed the case, see Fig 6 in the following section. In this regime, we have $|\mathcal{R}(G)| \sim d^3/6$, and the graph contains no far edges. Not surprisingly, there is little difference between SW1 and ER random graphs for large n.

Since the BA model is constructed such that it yield a fixed *average* vertex degree d, it should be compared to random graph models with the same vertex degree d instead of random graphs with a fixed edge drawing probabilities p. We have an asymptotically constant number of triangles for both ER and SW1: $\Delta_{\mathrm{ER}} \rightarrow d^3/6$ and $\Delta_{\mathrm{SW1}} \rightarrow d^3/6 - d + 2/3$, resp. Note that as a consequence the clustering coefficient vanishes asymptotically. In SW networks with *a priori* connectivity $k > 1$ we find of course a number of triangles that grows at least linearly with n, since the initial ($p = 0$) networks already contains $(k - 1)n$ triangles. The clustering coefficient stays finite for large n in this case (Watts, 1999).

The large vertex degree of the "early" vertices in the BA model suggests that there should be many more triangles than in ER or SW1 models. The expected degree of vertex s at "time" t is known (Dorogovtsev *et al.*, 2000): $d(s|t) = m[\sqrt{t/s} - 1]$. The probability of an edge between s and t, $t > s$, is therefore $p_{st} = md(s|t - 1)/2(t - 1)m$, where $2(t - 1)m$ is the sum of the vertex degrees at "time" $t - 1$. We have therefore

$$
\begin{aligned}
\langle \Delta \rangle &= \sum_{r<s<t} p_{rs}p_{st}p_{rt} \\
&\approx \frac{m^3}{8} \int_{1<r<s<t}^{n} (1/st^2) \left(\sqrt{\frac{s}{r}} - 1\right)\left(\sqrt{\frac{t}{r}} - 1\right)\left(\sqrt{\frac{t}{s}} - 1\right) \qquad (4.2) \\
&\sim Cm^3 \ln^3 n + \mathcal{O}(\ln^2 n)
\end{aligned}
$$

The l.h.s. panel in Fig. 5 shows Δ for typical BA-random graphs with $m = 2, \ldots, 8$ as a function of "time". The behavior of Δ in a individual growing network is well represented by equ.(4.2).

An extension of the BA model generates graph with $2 < \gamma \leq 3$. In addition to

the growth of the network, the model includes to rewiring operations: (i) addition of m new edges such that the initial points of the edges are chosen randomly while the terminal points are selected according to equ.(1.1), and (ii) rewiring of m randomly selected edges by leaving on endpoint fixed and re-attaching the other endpoint according to equ.(1.1). Since the scaling exponents depend on the relative frequency of the two rewiring operations, a quantitative comparison of chemical reaction networks with the extended BA model does not seem to meaningful at this point.

There is, however, a universal scaling relation between $P(d) \sim d^{-\gamma}$ and the degree $d(s|t) \sim (t/s)^{\beta}$ of vertex s and time t (Dorogovtsev et $al.$, 2000), namely $\beta = 1/(\gamma-1)$ and $2 \le \gamma \le 3$, i.e., $\frac{1}{2} \le \beta \ll 1$. Using the same reasoning as above the number of triangles should scale as $\langle \Delta \rangle \sim C(\beta)n^{2\beta-1} \ln n$ for $2 < \gamma < 3$. Thus we again expect the fraction of triangles to vertices to approach zero for large systems. The number of triangles in graphs with the same number of edges in vertices, on the other hand, increases with decreasing values of γ.

5. Longer Cycles

Much less can be said in general about longer relevant cycles. Computationally we find that the number $L = |\mathcal{R}| - \Delta$ of non-trivial relevant cycles has its maximum around $|E| \approx 0.74n^{3/2}$ independent of the random graph model, Fig 6. The scaling of is consistent with $L \sim Cn^{5/2}$, where the constant $C \approx 0.036$ is the same for ER and SW1 random graphs and $C \approx 0.016$ for the BA models. For small vertex degrees, $d \ll |V|^{1/2}$ we find $\mathcal{R}(G) \approx \nu(G)$, i.e., the MCB is (almost) unique.
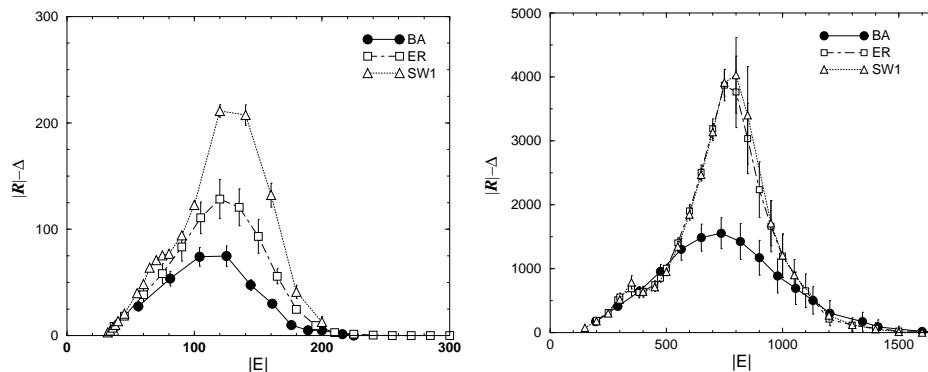

Figure 6: Relevant non-triangles in ER ($\square$), SW1 ($\triangle$), and BA ($\bullet$) random graphs with $n = 30$ (l.h.s) and $n = 100$ (r.h.s).

The cyclomatic number of a BA random graph is $\nu(G) \sim (m/2 - 1)n$; Hence, asymptotically, almost all relevant cycles must be long for $\beta < 1$, i.e., $\gamma > 2$. The l.h.s. of Fig. 5 shows that the average length of a relevant cycle grows logarithmically with n in the BA model. Not surprisingly, the slopes decrease with m.

6. Chemical Reaction Networks

6.1. A Metabolic Network

Metabolic networks form a particular class of chemical reaction networks which is distinguished by the fact that all reactions are associated with specific enzymes that catalyze the reaction.

For our analysis of metabolic graphs, we use the substrate graph of the `Ecoli1` core metabolism, a set of chemical reactions representing the central routes of energy metabolism and small-molecule building block synthesis. Similar to (Wagner and Fell, 2000), we omit the following substrates from the graph: CO_2, NH_3, SO_4, AMP, ADP, and ATP, their deoxy-derivatives, both the oxidized and reduced form of thioredoxine, organic phosphate and pyrophosphate. The resulting graph has $n = |V| = 272$ vertices and $|E| = 652$ edges. Its analysis is summarized in Table 1.

Recent results by Barabasi et al. (Jeong *et al.*, 2000) show that the degree distribution of a variety of metabolic networks follows a power law with scaling exponent $\gamma \approx 2.2$. Note that these author did not use the substrate graph Σ. Instead, they used the digraph representation of the reaction network, discussed in Section 3 and Figure 4, whose vertices are the substrates, the reactions, and the enzymes catalyzing the reaction. The numerical values of γ are not necessarily comparable between different graphical representations of reaction network.

The extended BA model (Albert and Barabasi, 2000), which is based on both growth and partial re-wiring of the networks can explain scaling exponents γ between 2 and 3. The discussion in (Fell and Wagner, 2000; Wagner and Fell, 2000) shows that a sequentially growing metabolic network is consistent with data because the evolutionary oldest metabolites have the largest vertex degrees.
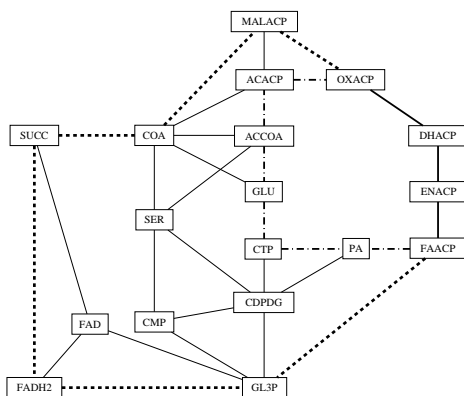

Figure 7: The subgraph of `Ecoli` spanned by the relevant cycles of length 9. Two of these long cycles are highlighted. The edges shown in bold are part of each of the 16 relevant 9-cycles.

The longest relevant cycles in a metabolic network are of particular interest since they reflect parts of the network that cannot easily be replaced by alternative routes. In Fig. 7 we show the largest such cycle in `Ecoli1`. We emphasize
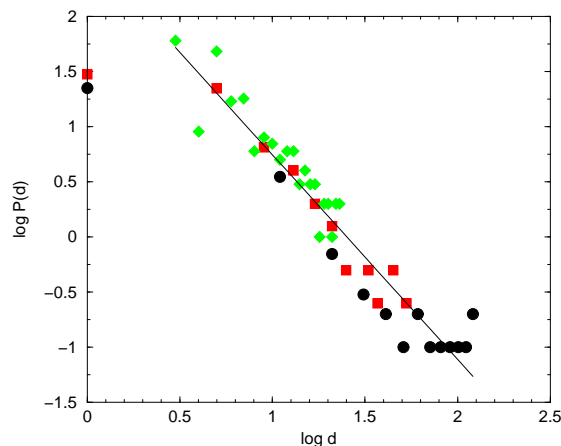
Figure 8: Degree distribution of the atmospheric reaction network of Earth. The symbols correspond to three different bin widths of the histogram for $P(d)$. The data are consistent with a power law with $\gamma = 1.86 \pm 0.09$ (full line).

that the cycles in our analysis represent routes for transmission of perturbations, but not necessarily of mass, as it is commonly considered in MFA. This is apparent from Fig.7, which does not correspond to a pathway from a biochemical chart, but links several pathways together. Note that the notion of a pathway requires from the outset the distinction between "substrates" or products and intermediates; for our purposes such a distinction is not necessary.

6.2. Planetary Atmospheres

It seems interesting to compare metabolic networks to reaction networks that are not governed by the enzymatic reactions. A class of large and well-understood models are the chemical networks of planetary atmospheres. The data reported here are taken from the book (Yung and DeMore, 1999). For details on these reaction networks we refer to (Yung and DeMore, 1999) and the references therein.

The largest network included in this study is a model of Earth's atmosphere which contains a large number of reactions involving halogen species including the CFCs implicated in global warming.

The atmospheres of the Jovian planets Jupiter, Saturn, Uranus, and Neptune are dominantly reducing. The thermodynamically stable form of carbon in the giant planets is methane CH_4. The photolysis of CH_4 leads to the production of higher hydrocarbons, some of which have been detected Earth-based or spacecraft observations. The network of the most important reactions inter-converting carbon species is denoted HC in Table 1 below.

Smaller networks model the atmospheres of the planets Mars and Venus, the Jovian satellite Io and the Saturn satellite Titan. The bulk of the atmospheres of both Mars and Venus is CO_2. While a pure CO_2 atmosphere should contain sizeable amounts of CO and O_2 small amounts of H_2O stabilize CO_2 through a network of reactions involving $\cdot OH$ radicals. In addition, both atmospheres contain N_2 and exhibit the associated chemistry of nitrogen oxides. Venus furthermore

exhibits an interesting sulfur chemistry. Io's thin atmosphere is dominated by the photo-chemistry of SO_2. Titan possesses a mildly reducing atmosphere exhibiting a rich hydrocarbon and nitrogen chemistry with HCN as a core species.

Fig. 8 shows that the atmosphere chemistry models also appear to have a scale free degree distribution with a scaling exponent $\gamma \approx 1.9$. This is surprising since these reaction networks could not have arisen by a stepwise mechanism. A possible explanation is a strong bias in the choice of chemical species and reaction pathways: the network models have been constructed to describe inter-conversion of a relative small number of dominating (or interesting) species, which naturally favors a "hub and spine" arrangement.

6.3. Comparison

Table 1 shows that the three random models BA, SW1, and ER agree at least qualitatively with each other. The BA random graphs exhibit a much broader distribution of cycle sizes (not shown) than the ER and SW1 models. As a consequence, the average cycle numbers for ER and SW1 have statistical uncertainty of about 2%, while the uncertainty of the BA values is 5 to 10 times higher. Note that ER and SW1 have a similar number of relevant cycles, but the cycles are slightly longer in SW1.

The substrate graphs of the planetary atmosphere models have a much larger average vertex degree. This accounts for the increased number of triangles and the lack of long relevant cycles.

Two features distinguish the metabolic network Ecoli1 from all three random network models:

(1) The number Δ of triangles is almost 10 times larger than expected. This can be explained by two effects. In part this might be an artifact of the substrate graph representation. The ratio $282/379 \approx 0.744$ indicates that almost all triangles are contained in 4-cliques, since in each 4-clique we have three triangles that belong to a particular MCB, while the fourth face of the tetrahedron is their $\oplus$-sum (Gleiss *et al.*, 2000).
   More importantly, however, the discussion in section 4 leads us to expect an increased number of triangles in scale free networks with small scaling exponent $\gamma < 3$, as is the case in metabolic networks (Wagner and Fell, 2000; Jeong *et al.*, 2000). A quantitative comparison between metabolic networks and the extended scale-free model (Albert and Barabasi, 2000) does not appear to be useful since the rewiring mechanism of the extended BA model is too artificial to apply to metabolic networks.
(2) There is a much smaller number of relevant pentagons and hexagons, which results in an overall somewhat reduced number of relevant cycles: 723 compared to about 1060 (BA), 904 (ER), and 805 (SW1). This is most likely again a consequence of the small value of the scaling exponent γ.

The atmosphere chemistry networks have a significantly larger average vertex degree. This explains the fact that almost all relevant cycles are triangles.

The vertices with the largest degree d in the raw data of many of the above networks are in some cases exceptional. In metabolic networks, for instance, ATP

Table 1: Cycle Structure of Networks.

Model	$|V|$	$|E|$	$|C|$	3	4	5	6	7	8	9	$\sum$
BA	272	625	MCB	78	158	124	20	0.4	0.01	0	380
			$\mathcal{R}$	81	285	527	161	5.5	0.4	0	1060
			$\mathcal{S}$	81	273	414	144	5.5	0.4	0	918
ER	272	625	MCB	18	58	163	131	11	0.4	0	381
			$\mathcal{R}$	18	61	212	528	82	3.2	0	904
			$\mathcal{S}$	18	61	205	311	68	3.2	0	666
SW1	272	625	MCB	15	46	131	167	21	1.1	0.03	381
			$\mathcal{R}$	15	48	157	427	151	7.1	0.2	805
			$\mathcal{S}$	15	48	155	301	108	6.5	0.2	634
Ecoli1	272	652	MCB	282	51	19	20	3	5	1	381
			$\mathcal{R}$	379	114	90	83	5	36	16	723
			$\mathcal{S}$	379	56	24	42	2	14	16	533
Earth	250	780	MCB	431	98	2	0	0	0	0	531
			$\mathcal{R}$	918	332	6	0	0	0	0	1256
			$\mathcal{S}$	918	303	6	0	0	0	0	1227
Titan	56	305	MCB	243	2	0	0	0	0	0	245
			$\mathcal{R}$	646	52	0	0	0	0	0	697
			$\mathcal{S}$	646	0	0	0	0	0	0	646
Venus	43	207	MCB	159	3	0	0	0	0	0	162
			$\mathcal{R}$	438	10	0	0	0	0	0	448
			$\mathcal{S}$	438	10	0	0	0	0	0	448
Mars	32	167	MCB	130	0	0	0	0	0	0	130
			$\mathcal{R}$	342	0	0	0	0	0	0	342
			$\mathcal{S}$	342	0	0	0	0	0	0	342
HC	40	299	MCB	260	0	0	0	0	0	0	260
			$\mathcal{R}$	1017	0	0	0	0	0	0	1017
			$\mathcal{S}$	1017	0	0	0	0	0	0	1017

is involved as "universal energy currency". Many of the reactions in planetary athmosphere involve a background gas atom as a means to removing excess energy from a reaction or photons $h\nu$. Following (Wagner and Fell, 2000) we argue that one should consider the network topology without these "special purpose" vertices. Almost all relevant cycles involving these exceptional species are triangles. We remark that their inclusion does not lead to qualitative changes of either the degree distribution or the distibution of relevant cycles apart from the obvious increase in the total number of cycles.

7. Discussion

We have shown that the union of minimal cycle bases, i.e., the set of *relevant cycles* forms a suitable framework for the investigation of the detours and, equivalently, cycles, in large network graphs. In the case of chemical reaction networks this view is reinforced by the close relationships of the cycle space and the sta-

tionary flux vectors of the network. We have also shown that the cycle structure of directed graphs can be discussed in terms of the underlying undirected graph. Since we have been interested in the propagation of perturbations through the chemical networks, undirected graphs are the appropriate model.

In metabolic flux analysis one is mostly concerned with the propagation of mass through the network. In this case directionality is crucial and directed graph models are required. In particularly, one would have to consider *directed circuit bases*, i.e., basis vectors with only non-negative entries. The discussion of sections 2 can indeed be generalized at least in part to directed cycle bases. In particular, all strongly connected networks have a directed cycle basis (Berge, 1985; Hartmann *et al.*, 1995). This topic will be discussed in more detail elsewhere.

The comparison of the metabolic networks and the reaction networks in planetary atmospheres show that the atmosphere models have a significantly higher vertex degree. This may be related to the fact that reactions in metabolic networks are mediated by specific enzymes; it is plausible that not all reactions that are chemically feasible have enzymes that make them part of the metabolism.

In (Wagner and Fell, 2000) a scaling exponent $\gamma = 1.59 \pm 0.21$ is reported for the substrate graph of the E. coli core metabolism, while we find $\gamma = 1.86 \pm 0.08$ for the atmosphere model of Earth. It is unclear at this point whether this difference is significant and whether we have here in fact a scaling exponent $\gamma < 2$. It is surprising that the atmosphere models exhibit a scale-free degree distribution because they most likely did not arise through stepwise addition of chemical species. More likely, the small-world geometry is an artifact of the construction of the models around a few dominating or interesting molecules which form the hubs of the network.

On the other hand, it appears reasonable to assume that the complex metabolic networks of present day organism have evolved from much simpler stages by stepwise addition of new metabolites. A BA-like mechanism of deriving new metabolites more readily from old, highly connected nodes appears plausible: the new metabolite is produced by a new enzyme which arose by mutation from an existing one – hence the nodes with the highest connectivity are most likely to give rise to variant enzymes which are likely to work with the same or a similar substrate. This view is consistent with the data (Fell and Wagner, 2000; Wagner and Fell, 2000): The most highly connected substrates are those that are identified as belonging to the oldest core of the metabolism (Morowitz, 1999).

Present-day metabolic networks have, however, a much larger than expected number of triangles and a correspondingly flatter degree distribution compared to the BA model. Strictly speaking, we do not know the biological significance of the relative paucity of longer cycles in metabolic networks. Nevertheless, we would like to venture a speculation.

Organisms are constantly exposed to environmental fluctuations requiring transitions in metabolic states. That is, a metabolic network needs to produce different outputs depending on the environment. Environments may vary rapidly, requiring rapid transition between metabolic states. Quite likely, networks with long cycles have longer transition times, because environmental perturbations may lead to prolonged oscillations in such networks. The dynamical system representation of metabolic networks required to test this idea rigorously lies beyond the scope of this article.

The extended BA model explains the decrease in γ by random rewiring events that are biased in such a way that the modified reactions are more likely to inter-convert two highly connected substrates. Such a mechanism appears unlikely since chemistry imposes very restrictive constraints on the vertices (chemical species) that can be connected (inter-converted) by a reaction. The loss of enzymes and hence edges in the network graph, on the other hand, appears to occur rather frequently: (Huynen *et al.*, 1999) for instance show that many organism use only parts of the citric-acid cycle. There does not seem to be a good reason why the loss of enzymes should preferentially affect poorly connected substrates, i.e., those that cannot be produced easily through alternate routes. We therefore argue that a degree-dependent rewiring process is not a likely explanation of the small values of γ. A selective advantage of rapid response, i.e., of small cycles, however, could well be the driving force behind this bias and explain the small γ-values of metabolic networks.

Acknowledgments

References

Albert, R. and Barabasi, A. L. Topology of evolving networks: local events and universality. *Phys. Rev. Lett.* pp. 5234–5237 (2000).

Amaral, L. A. N., Scala, A., Barthélémy, M., and Stanley, H. E. Classes of small world networks. *Proc. Natl. Acad. Sci. USA* **97**, 11149–11152 (2000).

Balaban, A. T., Fărcaşiu, D. and Bănică, R. Graphs of multiple 1,2-shifts in carbonium ions and related systems. *Rev. Roum. Chem.* **11**, 1205–1227 (1966).

Barabási, A.-L. and Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).

Barabási, A.-L., Albert, R. and Jeong, H. Mean-field theory for scale-free random networks. *Physica A* **173-187**, 272 (1999).

Barrat, A. and Weigt, M. On the properties of small-world network models. *Europ. Phys. J. B* **13**, 547– (2000).

Berge, C. *Graphs*. Amsterdam, NL: North-Holland (1985).

Bollobás, B. *Random Graphs*. London UK: Academic Press (1985).

Bollobás, B. *Modern Graph Theory*. New York: Springer (1998).

Chen, W.-K. On vector spaces associated with a graph. *SIAM J. Appl. Math.* **20**, 525–529 (1971).

Chickering, D. M., Geiger, D. and Heckerman, D. On finding a cycle basis with a shortest maximal cycle. *Inform. Processing Let.* **54**, 55–58 (1994).

Clarke, B. L. Stoichiometric network analysis. *Cell Biophys.* **12**, 237–253 (1988).

Dorogovtsev, S. N., Mendes, J. F. F. and Samukhin, A. N. Structure of growing networks with preferential linking. *Phys. Rev. Lett.* **85**, 4633–4636 (2000).

Edwards, J. D. and Palsson, B. Ø. The *escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. USA* **97**, 5528–5533 (2000).

Erdős, P. and Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci., Ser. A* **5**, 17–61 (1960).

Fell, D. A. *Understanding the Control of Metabolism*. London: Portland Press (1997).

Fell, D. A. and Wagner, A. Structural properties of metabolic networks: implications for evolution and modelling of metabolism. In: *Animating the cellular map* (Hofmeyr, J.-H. S., Rohwer, J. M. and Snoep, J. L., eds.), pp. 79–85. Stellenbosch, ZA: Stellenbosch University Press (2000).

Gleiss, P. M., Leydold, J. and Stadler, P. F. Interchangeability of relevant cycles in graphs. *Elec. J. Comb.* **7**, R16 [16pages] (2000).

Gleiss, P. M. and Stadler, P. F. *Relevant Cycles in Biopolymers and Random Graphs*. Tech. Rep. 99-07-042, Santa Fe Institute, Santa Fe, NM (1999). Presented at the Fourth Slovene International Conference in Graph Theory, Bled, Slovenia.

Hartmann, M., Schneider, H. and Schneider, M. H. Integral bases and *p*-twisted digraphs. *Europ. J. Combinatorics* **16**, 357–369 (1995).

Hartvigsen, D. Minimum path bases. *J. Algorithms* **15**, 125–142 (1993).

Hartvigsen, D. and Mardon, R. When do short cycles generate the cycle space. *J. Comb. Theory, Ser. B* **57**, 88–99 (1993).

Heinrich, R. and Schuster, S. *The Regulation of Cellular Systems*. New York: Chapman & Hall (1996).

Herzel, H. How to quantify "small world networks"? *Fractals* **6**, 301–303 (1998).

Horton, J. D. A polynomial-time algorithm to find the shortest cycle basis of a graph. *SIAM J. Comput.* **16**, 359–366 (1987).

Huynen, M. A., Dandekar, T. and Bork, P. Variation and evolution of the citric-acid cycle: a genomic perspective. *Trends Microbiol* **7**, 281–291 (1999).

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. and Barabasi, A. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).

Kammermeier, S., Neumann, H., Hampel, F. and Herges, R. Diels-Alder reactions of tetradehydrodianthracene with electron-rich dienes. *Liebigs Ann.* **1996**, 1795–1800 (1996).

Leydold, J. and Stadler, P. F. Minimal cycle basis of outerplanar graphs. *Elec. J. Comb.* **5**, R16 (1998). See http://www.combinatorics.org and Santa Fe Institute Preprint 98-01-011.

McKee, T. A. Induced cycle structure and outerplanarity. *Discr. Math.* **223**, 387–392 (2000).

Montoya, J. M. and Solé, R. V. *Small World Patterns in Food Webs*. Tech. Rep. 00-10-059, Santa Fe Institute (2000).

Morowitz, H. J. A theory of biochemical organization, metabolic pathways, and evolution. *Complexity* **4** (6), 39–53 (1999).

Newman, M. E. J., Moore, C. and Watts, D. J. Mean-field solution of the small-world network model. *Phys. Rev. Lett.* **84**, 3201–3204 (2000).

Newman, M. E. J. and Watts, D. J. Renormalization group analysis of the small-world network model. *Phys. Lett. A* **263**, 341–346 (1999).

Pandit, S. A. and Amritkar, R. E. Characterization and control of small-world networks. *Phys. Rev. E* **60**, R1119–R1122 (1999). Chao-dyn/9901017.

Schilling, C. H., Letscher, D. and Palsson, B. Ø. Theory for the systematic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.* **203**, 229–248 (2000).

Schuster, S., Fell, D. A. and Dandekar, T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnol.* **18**, 326–332 (2000).

Sen, A. K. Quantitative analysis of metabolic regulation. A graph-theoretic approach using spanning trees. *Biochem. J.* **275**, 253–258 (1991).

Stadler, P. F. Minimal cycle bases of Halin graphs. *J. Graph Theory* (2000). Submitted.

Stepanec, G. F. Basis systems of vector cycles with extremal properties in graphs. *Uspekhi Mat. Nauk. 2* **19**, 171–175 (1964). (Russian).

Temkin, O. N., Zeigarnik, A. V. and Bonchev, D. *Chemical Reaction Networks: A Graph-Theoretical Approach*. Bocal Raton, FL: CRC Press (1996).

Vismara, P. Union of all the minimum cycle bases of a graph. *Electr. J. Comb.* **4**, 73–87 (1997).

Wagner, A. and Fell, D. A. *The small world inside large metabolic networks*. Tech. Rep. 00-07-041, Santa Fe Institute (2000).

Watts, D. J. *Small Worlds*. Princeton NJ: Princeton University Press (1999).

Watts, D. J. and Strogatz, H. S. Collective dynamics of "small-world" networks. *Nature* **393**, 440–442 (1998).

Yung, Y. L. and DeMore, W. B. *Photochemistry of Planetary Athmospheres*. New York: Oxford University Press (1999).

Zeigarnik, A. V. On hypercycles and hypercircuits in hypergraphs. In: *Discrete Mathematical Chemistry* (Hansen, P., Fowler, P. W. and Zheng, M., eds.), vol. 51 of *DIMACS series in discrete mathematics and theoretical computer science*, pp. 377–383. Providence, RI: American Mathematical Society (2000).

Zykov, A. A. *Theory of Finite Graphs*. Novosibirsk, USSR: Nauka (1969). (Russian).