Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 11, Number 1 (2015), pp. 241–256

http://www.ripublication.com/gjpam.htm

Loop Saturated Graphs and Index of Assembly Words

Tilahun A. Muche
Savannah State University,
Department of Mathematics,
Savannah State University, Savannah, GA, 31404 USA.
E-mail: muchet@savannahstate.edu, tilahun_rach@yahoo.com

Abstract

The assembly number of an *assembly graph* increases if the polygonal paths are forced to travel along certain edges. This enforcement can be obtained by introducing loops on the edges and hence introducing the necessity for the polygonal paths to visit vertices of the loops [9]. In this paper I give formulae for the assembly number of graphs obtained by adding loops on edges, and investigate the sizes of their assembly numbers and corresponding rigid vertices. I also investigate the odd index of loop saturated graphs.

AMS subject classification:
Keywords: Loop saturated graphs, Assembly numbers, Odd index, Sequences.

1. Introduction

Gene assembly is a biological process that takes place in a large group of unicellular organisms called ciliates. Ciliates have two distinct types of nuclei called micronuclei (MIC) and macronuclei (MAC) [1, 4]. Gene assembly occurs during sexual reproduction of certain species of ciliates, and this process transforms the micronucleus into the macronucleus through a number of splicing operations. Spatial graphs with 4-valent rigid vertices and two endpoints with single valency, called assembly graphs, model DNA recombination processes that appear in certain species of ciliates. Recombined genes are modeled by certain types of paths in an assembly graph that make a "perpendicular" turn at each 4-valent vertex of the graph called polygonal paths. The assembly number of an assembly graph is the minimum number of polygonal paths that visit each vertex exactly once [5, 9].

An assembly graph $\hat{\Gamma}$ obtained from a given assembly graph Γ by substituting every edge of Γ by a loop is called a loop-saturated graph.

2. Assembly Words and Assembly Graphs

A graph $G = (V, E)$ is a pair consisting of a set of vertices V and a set of edges E where the two endpoints of an edge in E are vertices in V. If e is an edge and v is an end point of e, then e is said to be *incident* to v. The number of edges incident to a vertex v is called the *degree* of v. A *loop* is defined as an edge with one endpoint, contributes 2 to the degree of a vertex [9, 2, 3].

A *4-valent rigid vertex* is a vertex of degree 4 for which a cyclic order of edges is specified. For a 4-valent rigid vertex v, if its incident edges appear in order e_1, e_2,e_3, e_4, we say that e_2 and e_4 are neighbors with respect to v to e_1 (or e_3). Vice versa, e_1 and e_3 are neighbors to e_2 (or e_4).

Definition 2.1. An *assembly graph* is a finite connected graph in which all vertices are rigid and have degree 1 or 4. A vertex with degree 1 is called an *endpoint*.

The number of 4-valent vertices in an assembly graph Γ is called the size of Γ and is denoted by $|\Gamma|$.The assembly graph is called trivial if $|\Gamma| = 0$.Two types of paths are of interest: (a) paths in which consecutive edges are never neighbors with respect to their common incident vertex and (b) paths in which every pair of consecutive edges are neighbors with respect to their common incident vertex. A path of type (a) where no edge is repeated is called a *transverse* path, or simply a *transversal*. A path of type (b) where no vertex is repeated is called a *polygonal* path. Graphs that have an Eulerian transversal are called *simple assembly graphs*. We note that in a simple assembly graph, if a vertex v is an endpoint of a loop e, then e must be a neighbor of itself [9].

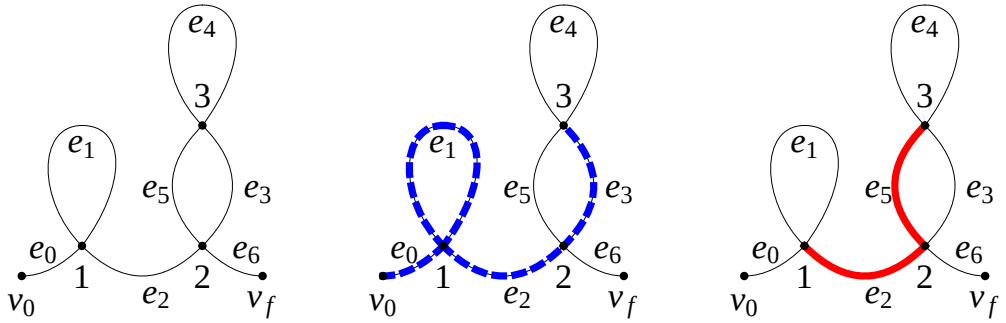

Figure 1: A transverse path $(v_0, e_0, 1, e_1, 1, e_2, 2, e_3, 3)$(b), and a polygonal path $(1, e_2, 2, e_5, 3)$ (c) for the simple assembly graph (a).

Definition 2.2. An assembly graph Γ is called *simple* if there is a transverse Eulerian path in Γ; otherwise, it is called *non simple assembly graph.*

In a simple assembly graph Γ there is a transverse path γ that contains every edge

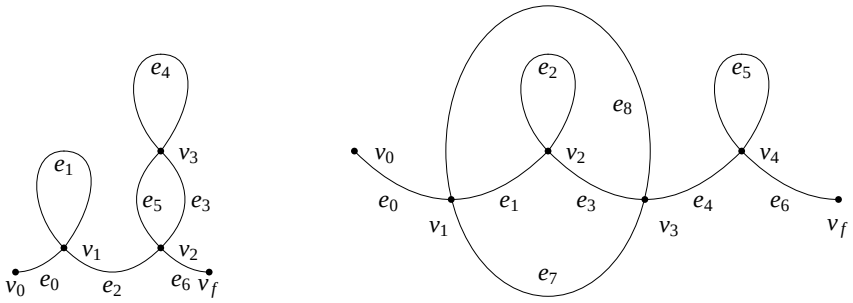

Figure 2: A simple assembly graph that corresponds to an assembly word $w = 112332$ (left) and a non-simple assembly graph (right).

exactly once. An example of a simple assembly graph and a non-simple assembly graph are depicted in Figure 2(left) and Figure 2 (right), respectively.

Assembly graphs are of particular interest because of their recent use to model genome rearrangement in ciliates. Ciliates are unicellular organisms which contain two types of nuclei, the germline (micronuclear) and somatic (macronuclear). The micronucleus contains segments of DNA found in the macronucleus but often in a permuted order and separated by non-coding DNA. During sexual reproduction the micronuclear genome undergoes massive elimination of non-coding DNA and rearrangement to obtain a new macronucleus. Both transversals and polygonal paths model specific parts of the DNA rearrangement phe- nomenon. In particular, the transversal represents the micronuclear DNA segment prior to the recombination, the rigid vertices indicate the recombination sites, and polygonal paths are DNA segments after the rearrangements [9].

I refer the reader to [5] for a more thorough treatment of the biological background and I recommend [3] for more on the assembly graph model.

Given a simple assembly graph Γ, designate one of the endpoints as initial (i) and the other endpoint as terminal (t). I call such Γ an *oriented* (or *directed*) simple assembly graph with direction from i to t. I consider the transverse path of a directed simple assembly graph as a path starting at the vertex i and terminating at the vertex t.

A double occurrence word w is a word containing symbols (or letters) from a finite alphabet such that every symbol in w appears exactly twice. We now establish a convention for representing simple assembly graphs by words. Let Γ be a simple assembly graph with vertices $v_1, v_2, v_3, \ldots, v_n$. Given a transverse Eulerian path of Γ by $\gamma = (v_{i_0}, e_1, v_{i_1}, \ldots, e_{2n+1}, v_{i(2n+1)})$ for $i_k \in \{1, \ldots, n\}$, note that all vertices except endpoints, v_{i_0} and $v_{i_{2n+1}}$, are visited exactly twice. Thus, we can represent Γ by the double occurrence word $v_{i_1}, v_{i_2}, \ldots, v_{i_{2n}}$ [6, 8].

A double occurrence word w with n distinct symbols has size n and length $|w| = 2n$.

3. Assembly Number and Loop Saturated Graphs

Two paths are *disjoint* if they do not have a vertex in common. I are interested in disjoint polygonal paths that visit every vertex in an assembly graph. A pairwise disjoint set $\{\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_n\}$ of polygonal paths in Γ is called *Hamiltonian* if their union contains all 4-valent vertices of Γ. In particular the set of vertices $V(\Gamma)$ is a Hamiltonian set of singletons. A polygonal path γ with no repeating vertices is called Hamiltonian if the set γ is Hamiltonian.

3.1. Assembly Number

Let Γ be a non-trivial assembly graph. The *assembly number* of Γ that is denoted by $An(\Gamma)$ is defined by $An(\Gamma)$= min{ k : there exists a Hamiltonian set of polygonal paths $\{\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_n\} in\ \Gamma\}$. In particular the assembly number of a graph gives the minimal number of genes that can be encoded by a corresponding DNA sequence.

Note: There is only one assembly word with one letter $w = 11$. I denote the corresponding assembly graph by $\Gamma_{(1)}$, which is *the loop*. See the figure below.

Example 3.1. In Figure 3, an assembly graph is given with Hamiltonian set of polygonal paths $\gamma = \{\gamma_1, \gamma_2, \gamma_3\}$.

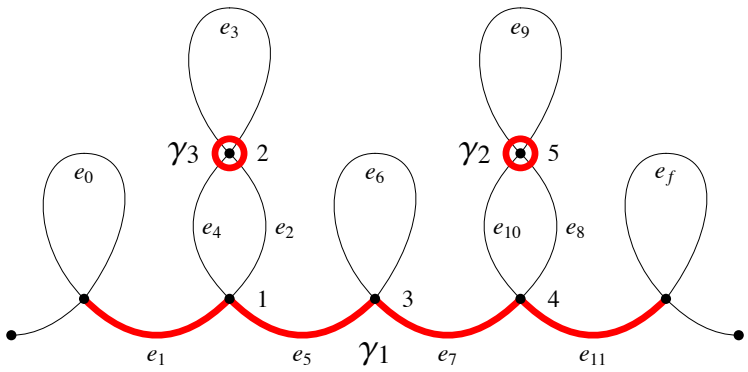

Figure 3: Hamiltonian set of polygonal paths $\gamma = \{\gamma_1, \gamma_2, \gamma_3\}$.

Definition 3.2. The assembly graph $\hat{\Gamma}$ obtained from a given assembly graph Γ with substituting every edge by a loop $\Gamma_{(1)}$ is called the *assembly graph obtained from* Γ *by loop-saturation*, or a *loop-saturated graph*.

The assembly graph $\hat{\Gamma}^{\circ}$ obtained from a given assembly graph Γ by substituting every edge by a loop $\Gamma_{(1)}$, except the edges incident to the endpoints, is called the *assembly graph obtained from* Γ *by interior loop-saturation*. A graph obtained in this manner is said to be an *interior loop-saturated graph*.

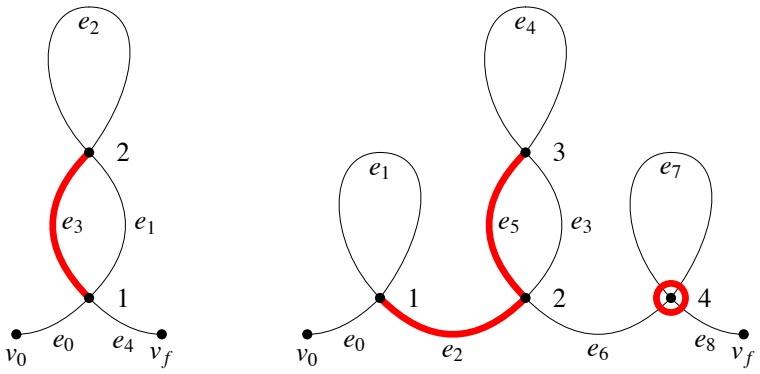

Figure 4: Simple assembly graphs and polygonal paths in red lines; in the right, there are two polygonal paths, one of which is a singleton containing the vertex 4.

Loop-saturated and interior loop-saturated graphs obtained from graphs with 1 and 2 vertices are depicted in Figure 5. If $\hat{\Gamma}$ is obtained from Γ by loop-saturation, then all vertices of Γ are vertices of $\hat{\Gamma}$. In fact, the vertices of $\hat{\Gamma}$ consist of those from Γ and vertices incident to loops added by loop-saturation. Moreover, any loop in $\hat{\Gamma}$ is incident to a vertex that is not in Γ. The loop $\Gamma_{(1)}$ is a loop-saturation of the trivial graph with two endpoints and no 4-valent vertices.

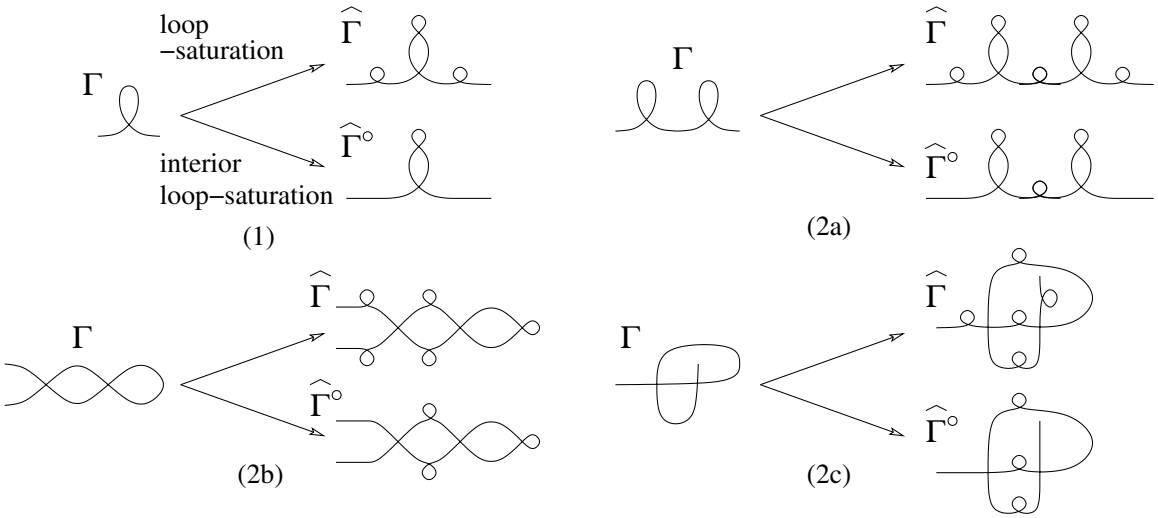

Figure 5: Loop-saturated and interior loop-saturated graphs.

Definition 3.3. For an integer $n \geq 0$, the set of assembly graphs obtained by interior loop-saturation from assembly graphs of size n is denoted by $\mathcal{G}_n$.

For example, $\mathcal{G}_0$ consists of only the trivial graph, $\mathcal{G}_1$ consists of only one graph corresponding to the assembly word 1221 depicted in Figure 5 top left. For $n = 2$, $\mathcal{G}_2$ is the set of assembly graphs corresponding to the assembly words 1221334554, 1223443551 and 1223441553. These graphs correspond to the interior loop-saturated graphs of size 2 in Figure 5 and are depicted at the bottom right of Figure 5(2a), (2b) and (2c), respectively. Note that the assembly number of these three graphs is 1.

Definition 3.4. The *length* of a polygonal path γ, denoted $|\gamma|$, is the number of 4-valent vertices that the path contains.

Definition 3.5. Let $\gamma = \{\gamma_1, \ldots, \gamma_k\}$ be a set of polygonal paths in an assembly graph Γ and assume that $|\gamma_1| \geq |\gamma_2| \geq \cdots \geq |\gamma_k|$. The *height* sequence for γ, denoted by $\mathrm{Ht}(\gamma)$, is a sequence of positive integers $(|\gamma_1|, \ldots, |\gamma_k|)$.

Lemma 3.6. Let H be a loop-saturated or interior loop-saturated assembly graph with $|H| > 2$, and $\gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_k\}$ be a minimal Hamiltonian set of polygonal paths. Then $|\gamma_i|$ is odd for each $i = 1, 2, \ldots, k$.

Proof. Let H be a loop-saturated or interior loop-saturated graph obtained from Γ, and let $L = V(H) \setminus V(\Gamma)$. Then, L is the set of all new vertices obtained by loop saturation of Γ. Note that the vertices in every polygonal path alternate between vertices in $V(\Gamma)$ and those in L. We show that each polygonal path γ_i in a minimum Hamiltonian set of polygonal paths starts and ends at a loop. This implies that $|\gamma_i|$ is odd.

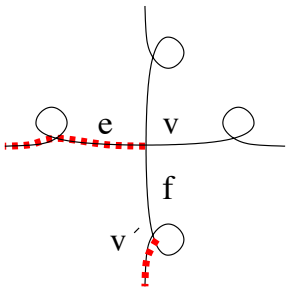

Figure 6: A polygonal path that doesn't end at a loop.

Suppose γ_i is a path in a Hamiltonian set of polygonal paths γ that ends at a vertex $v \in V(\Gamma)$ (see Figure 6). There are two cases; either v is incident to an endpoint and $H = \hat{\Gamma}^\circ$ is interior loop-saturated, or v is not incident to an endpoint. In the latter case, consider the end edge of γ_i denoted by e. Both edges that are neighbors of e with respect to v are incident to vertices in L (by construction). Let v' be one of these vertices and f the edge with endpoints v and v'. There exists a polygonal path γ_j that visits v' so v' must be an end-vertex of γ_j. Then substitute the two polygonal paths γ_i, γ_j in the Hamiltonian set of polygonal paths γ with a single path $\gamma_i f \gamma_j$. The new set of paths is

Hamiltonian and contains fewer paths than γ. Hence, γ is not a minimum Hamiltonian set of polygonal paths.

The case when $H = \hat{\Gamma}^{\circ}$ is interior loop-saturated and v is adjacent to an endpoint follows similarly. In this case, because H has more than two vertices, the vertex v is adjacent to at least two and at most three loops, one of which must be an endpoint of a neighbor of the last edge e of the polygonal path γ_i ending at v. ■

In the case when $|H| = 2$, H must be the interior loop-saturated graph of $\Gamma_{(1)}$, and it has a polygonal path of length 2.

Theorem 3.7. For any assembly graph Γ with $|\Gamma| = n$, the loop-saturated graph $\hat{\Gamma}$ obtained from Γ has assembly number $n + 1$ and the interior loop-saturated graph $\hat{\Gamma}^{\circ}$ has assembly number $n - 1$.

Proof. Let $\hat{\Gamma}$ be loop-saturated graph obtained from Γ with $|\Gamma| = n$. Because Γ has $2n + 1$ edges, $|\hat{\Gamma}| = 3n + 1$. We show that $\mathrm{An}(\hat{\Gamma}) = n + 1$. Let $\gamma = \{\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_s\}$ be a minimum Hamiltonian set of polygonal paths of $\hat{\Gamma}$ and set $|\gamma_i| = d_i$ for all $i = 1, \cdots, s$. Suppose m is the number of singletons in γ. The remaining $(s - m)$ of any polygonal paths visit all vertices of Γ because every path in γ starts and ends at a loop as shown in the proof of Lemma 3.6.

Let k_i denote the number of vertices in γ_i for $i = 1, 2, 3, \ldots, (s-m)$. Then $\sum_{i=1}^{s-m} k_i = n$. Hence

$$\begin{aligned} 3n + 1 = \sum_{i=1}^{s} d_i &= \sum_{i=1}^{s-m} d_i + \sum_{i=s-m+1}^{s} d_i \\ &= \sum_{i=1}^{s-m} (2k_i + 1) + m. \end{aligned}$$

This implies that $\sum_{i=1}^{s-m} (2k_i + 1) = 2\sum_{i=1}^{s-m} k_i + s - m = 3n + 1 - m$. Consequently $2n + s = 3n + 1$ and $s = n + 1 = \mathrm{An}(\hat{\Gamma})$.

The claim for $\hat{\Gamma}^{\circ}$ follows similarly since $|\hat{\Gamma}^{\circ}| = 3n - 1$. ■

Definition 3.8. For $n \geq 0$, we denote by $L^n(\Gamma)$ the loop-saturated simple assembly graph obtained from a simple assembly graph Γ of size n and by $L^n(\Gamma_{(1)})$ an n-th loop-iteration of $L^1(\Gamma_{(0)})$=$\Gamma_{(1)}$=$L^0(\Gamma_{(1)})$, a loop-saturated graph that is formed by successively loop-saturated of $\Gamma_{(0)}$, see Fig 7.

Theorem 3.9. The size of $L^n(\Gamma_{(1)})$ that is $|L^n(\Gamma_{(1)})| = \sum_{i=0}^{n} 3^i$ for $n \geq 0$ and $\mathrm{An}(L^n(\Gamma_{(1)})) = 3^n - |L^{n-1}(\Gamma_{(1)})|$, $n \geq 1$.

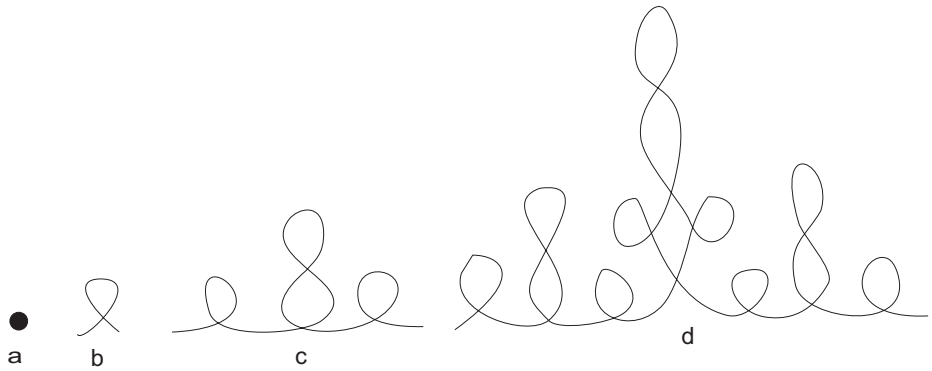

Figure 7: (a)=$\Gamma_{(0)}$,(b)=$L^1(\Gamma_{(0)})$=$\Gamma_{(1)}$, (c)=$L^2(\Gamma_{(0)})$=$L^1(\Gamma_{(1)})$,(d)=$L^3(\Gamma_{(0)})$=$L^2(\Gamma_{(1)})$.

Proof. $|L^0(\Gamma_{(1)})| = |\Gamma_{(1)}| = 1$ and $\text{An}(L^0(\Gamma_{(1)})) = 1$. The second loop-iteration of a simple assembly graph $L^1(\Gamma_{(1)})$ is obtained from $L^0(\Gamma_{(1)})$. By Lemma 3.7, $\text{An}(L^1(\Gamma_{(1)}) = |L^0(\Gamma_1)|+1 = 3^0+1$ and $|L^1(\Gamma_{(1)})| = 3\text{An}(L^1(\Gamma_{(1)})-2$. This implies $\text{An}(L^1(\Gamma_{(1)})) = 3^1 - 3^0$ and $|L^1(\Gamma_{(1)})| = 3^1 + 3^0 = \sum_{i=0}^{1} 3^i$. We use induction to prove the statement. It is true for $n = 0$. Assume that it is true for some positive integer $s > n$. Then

$$\begin{aligned} |L^s(\Gamma_{(1)})| = \sum_{i=0}^{s} 3^i. \text{ But } |L^{s+1}(\Gamma_{(1)})| \\ = 3|L^s(\Gamma_{(1)})| + 1 = 3\sum_{i=0}^{s} 3^i + 1 \\ = \sum_{i=0}^{s} 3^{i+1} + 1 = \sum_{i=0}^{s+1} 3^i. \end{aligned}$$

Next we show that $\text{An}(L^n(\Gamma_{(1)})) = 3^n - |L^{n-1}(\Gamma_{(1)})|$. It is true for $n = 1$ as $\text{An}(L^1(\Gamma_{(1)})) = 3^1 - |L^0(\Gamma_{(1)})| = 2$. Assume that the statement is true for some $t \geq n$. This implies $\text{An}(L^t(\Gamma_{(1)})) = 3^t - |L^{t-1}(\Gamma_1)|$. By Lemma 3.7, $\text{An}(L^{t+1}(\Gamma_{(1)})) = |L^t(\Gamma_{(1)})| + 1 = \sum_{i=0}^{t} 3^i + 1$, but

$$\sum_{i=0}^{t+1} 3^i = \sum_{i=0}^{t} 3^i + 3^{t+1}.$$

This implies

$$\sum_{i=0}^{t+1} 3^i - \sum_{i=0}^{t} 3^i = 3^{t+1}$$

and

$$\begin{aligned} 3^{t+1} - \sum_{i=0}^{t} 3^i &= \sum_{i=0}^{t+1} 3^i - 2\sum_{i=0}^{t} 3^i \\ &= \sum_{i=0}^{t} 3^i + 1 = |L^t(\Gamma_{(1)})| + 1 = \mathrm{An}(L^{t+1}(\Gamma_{(1)})). \end{aligned}$$

Consequently $\mathrm{An}(L^{t+1}(\Gamma_{(1)})) = 3^{t+1} - |L^t(\Gamma_{(1)}))|$. ■

Lemma 3.10. For each $n \geq 1$, $\mathrm{An}(L^n(\Gamma_{(1)})) = 3\mathrm{An}(L^{n-1}(\Gamma_{(1)})) - 1$.

Proof. We showed that $|L^n(\Gamma_{(1)})| = 3|L^{n-1}(\Gamma_{(1)})|+1$ and $\mathrm{An}(L^n(\Gamma_{(1)})) = |L^{n-1}(\Gamma_{(1)})|+1$. This implies

$$3\mathrm{An}(L^n(\Gamma_{(1)})) = 3|L^{n-1}(\Gamma_{(1)})| + 3 = |L^n(\Gamma_{(1)})| + 2$$

and

$$3\mathrm{An}(L^n(\Gamma_{(1)})) - 1 = |L^n(\Gamma_{(1)})| + 1.$$

Hence from Theorem 3.9, $\mathrm{An}(L^n(\Gamma_{(1)})) = 3\mathrm{An}(L^{n-1}(\Gamma_{(1)})) - 1$, for $n \geq 1$. ■

Properties: 1) Set $\beta(n) = |L^n(\Gamma)|$. Then $\beta(n)$=$3n + 1$ for $n \geq 0$.

Table 1: **The sequence of numbers that corresponds to the size of loop-saturated simple assembly graphs obtained from Γ of size $|\Gamma| = n$.**

$\lvert\Gamma\rvert = n$	$\lvert L^n(\Gamma)\rvert$
0	1
1	4
2	7
3	10
4	13
5	16
6	19
7	22
8	25
$OEIS$	A016777

Let $\Gamma_{(1)}$ be a loop-iterated assembly graph obtained from a trivial graph $\Gamma_{(0)}$ such that $|\Gamma_{(0)}|$=0.

2) Consider the sequence of numbers $\mathrm{An}(L^n(\Gamma_{(0)}))$. Set $\mathrm{An}(L^n(\Gamma_{(0)})) = \phi(n)$ and $\phi(0) = 0$.

$$\phi(n) = \begin{cases} \dfrac{3^{n-1}+1}{2} & \text{if } n \geq 1 \\ 0 & \text{if } n = 0 \end{cases}$$

Table 2: **The sequence of numbers that corresponds to the assembly number of the n-th loop-iterated simple assembly graphs obtained from $\Gamma_{(0)}$.**

Rigid vertices n	$\phi(n)$
0	0
1	1
2	2
3	5
4	14
5	41
6	122
7	365
8	1094
OEIS	A007051

3) Let $m(i) = \phi(i) + \phi(i+1)$ for $i \geq 1$.

$$m_{(i)} = \phi(i) + \phi(i+1) = \frac{3^{n-1}+1}{2} + \frac{3^n+1}{2} = 1 + 2 * 3^{n-1} \text{ for } n \geq 1.$$

4) Set $|L^n(\Gamma_{(0)})| = \psi(n)$.

$$\psi(n) = \frac{3^n - 1}{2} \text{ for } n \geq 0.$$

5) Define $\chi(n) = h(\dfrac{\psi(n)}{\psi(n+1)}) = \dfrac{2\psi(n) + \psi(n+1)}{\psi(n) + 2\psi(n+1)}$, for $n \geq 0$. Then $\chi(n) = \dfrac{3^{n+1}-1}{3^{n+1}+1}$.

The Tables indicate that these sequences are not part of the known sequences listed in [10].

Table 3: **The sequence of numbers obtained by adding two consecutive assembly numbers of loop-iterated saturated assembly graphs.**

$Rigid\ vertices\ n$	$m(n)$
1	3
2	7
3	19
4	55
5	163
6	487
7	1459
8	4375
$OEIS$	$A100702$

Table 4: **The sequence of numbers corresponds to the size of the n-th loop iterated saturated assembly graphs.**

$Rigid\ vertices\ n$	$\psi(n)$
0	0
1	1
2	4
3	13
4	40
5	121
6	364
7	1093
8	3280
$OEIS$	$A003462$

4. Loop Saturated Graphs and Index Number

Definition 4.1. Let w be a double occurrence word. A rigid vertex labeled letter m is said to be *odd* if the number of the letters that are neither 1 nor the last letter of w between two copies of m is an odd integer [11].

The number of odd letters is called the *odd index* of w and denoted by $I_{odd}(w)$. The same terms are defined corresponding assembly graphs, and the odd index of an assembly graph Γ is denoted by $odd(\Gamma)$ [7].

Example 4.2. For $w = 121323$, the vertices 1 and 3 are odd as $I_{odd}(1) = 1 = I_{odd}(3)$ but $I_{odd}(2) = 0$ is even, so that $I_{odd}(w) = I_{odd}(121323) = 2$.

Table 5: **This sequence corresponds to the function defined by** $h\left(\frac{a}{b}\right) = \frac{2a+b}{a+2b}$ **where** $\phi(n) = a$ **and** $b = \psi(n+1)$ **and** $lim_{n\to\infty}\psi(n) = 1$**.**

$Rigid\ vertices\ n$	$\chi(n)$
0	$\frac{1}{2}$
1	$\frac{4}{5}$
2	$\frac{13}{14}$
3	$\frac{40}{41}$
4	$\frac{121}{122}$
5	$\frac{364}{365}$
6	$\frac{1093}{1094}$
7	$\frac{3280}{3281}$
8	$\frac{9841}{9842}$
$OEIS$	$A001764$

Definition 4.3. The set of first k natural numbers is denoted by $\mathcal{N}_k = \{1, 2, 3, \ldots, k\}$. The set $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ is called an *alphabet* if for all $i \in \mathcal{N}_m$, α_i are arbitrary symbols. The elements of $\mathcal{A}$ are then the *letters* of the alphabet and $\mathcal{A}$ is m letter alphabet. If $x \in \mathcal{A}^n$, i.e., $x = x_1x_2\ldots x_n$ with each $x_i \in \mathcal{A}$, we say that x is a word of length n over the alphabet $\mathcal{A}$.

A *subword* of length k of the word $x = x_1x_2\ldots x_n$ is any word $x_sx_{s+1}\ldots x_{s+k-1}$ where $s \in N_{n-k+1}$ and $k \in \mathcal{N}_n$ [8].

Definition 4.4. Let w and w^* be double occurrence words that correspond to a simple assembly graph Γ_w and its loop-iterated saturated graph $L^1(\Gamma_w)$ respectively. Then we

write $p \in (w \cap w^*)$ if p is not a vertex labelled to a loop in $L^1(\Gamma_w)$. We call vertex p a *root vertex.*

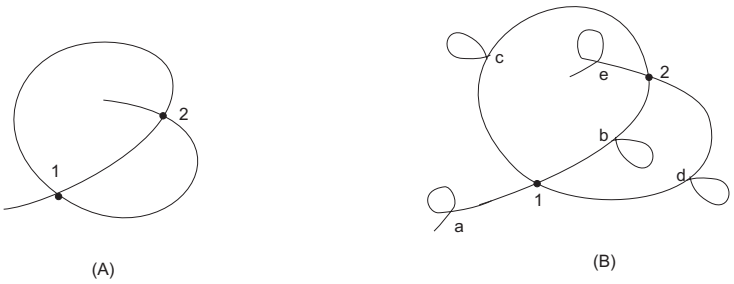

Figure 8: $(A) = \Gamma_w$, $w = 1212$, $(B) = L^1(\Gamma_w) = \Gamma_{w^*}$, $w^* = aa1bb2cc1dd2ee$.

Note: If p is among the letters in a double occurrence word w, we write $p \in w$; moreover, all letters $p \in w$ are root vertices with respect to the double occurrence word w^*.

Example 4.5. See Fig 8. The double occurrence words $w^* = aa1bb2\ cc1dd2ee$ and $w = 1212$. The root vertices are 1 and 2 as $1, 2 \in (w \cap w^*)$.

Definition 4.6. Let w be a double occurrence word with initial and terminal end points 1 and z respectively. If $a \in w$ such that $a_i = a = a_j$ for some $i, j \in \mathcal{N}$, then we write $I_{last}(a)$ is even and $I_{last}(a)$ is odd if the size of the end points between a_i and a_j are even and are odd respectively.

Example 4.7. Consider the double occurrence word $w = 12134234$. Then,
$I_{last}(2) = 2$, end points 1 and 4.
$I_{last}(3) = 1$, end point only 4.
$I_{odd}(2) = 1 = I_{odd}(3)$.

Remarks:

- No two of the first end points of a double occurrence word w appear between two identical symbols.
- if $I_{last}(a)$ is even, then either

  1) none of the end points appears between two copies of a i.e., $I_{last}(a) = 0$ or,
  2) 1 and z appear between two copies of a and $I_{last}(a) = 2$.

- if $I_{last}(a)$ is odd then $I_{last}(a) = 1$ in which case only 1 or only z appears.

Lemma 4.8. For $p \in w$ and a subword $pw'p$ of w, $I_{last}(p) + I_{odd}(p)$ equals the size of the subword w'.

Theorem 4.9. $p \in w^*$ and $I_{odd}(p)$ is odd if and only if $p \in w$ and for a subword $pw'p$ of w the size of the subword w', $|w'|$, is odd.

Proof. Suppose $p \in w^*$ and $I_{odd}(p)$ is odd. Then p can not be a vertex that corresponds to a loop because pp is a subword of w^* and $I_{odd}(p)$ is even. Hence $p \in (w \cap w^*)$ and is a root vertex.

Consider a subword $pw'p = pp_1p_2 \dots p_{j+1}p$ of w. For some j, there exist a sequence of alphabets $a_1, a_2, \dots, a_j$ labeled to loops of a simple assembly graph Γ_{w^*} that corresponds to a double occurrence word w^* such that $pyp = pa_1a_1p_1a_2a_2p_2a_3a_3, \dots, a_j a_j p_{j+1}p$ is a subword of w^*. Because $I_{last}(y) = I_{last}(p) = 0$, $I_{odd}(p)$ is odd implies, $I_{last}(p) + I_{odd}(p) = |y| = |a_1a_1p_1a_2a_2p_2a_3\ a_3 \dots a_ja_j\ p_{j+1}| = |p_1p_2, \dots, p_{j+1}| + 2|a_1a_2, \dots, a_j| = 2n + 1$ for some $n \in \mathcal{N}$. Hence w' is a subword of w and $|w'| = |p_1p_2 \dots p_{j+1}| = 2(n - j) + 1$ is odd.

Conversely, suppose $p \in w$ and $pw'p$ is a subword of w with $|w'|$ is odd. There is a subword $pa_ia_{i+1} \dots a_{i+m}p$ of w such that $|w'| = |a_ia_{i+1} \dots a_{i+m}| = 2k + 1$ for some $k \in \mathcal{N}$. The vertices $p, a_i, a_{i+1}, \dots, a_{i+m} \in (w \cap w^*)$. This implies a subword $t = px_jx_ja_ix_{j+1}x_{j+1}a_{i+1} \dots x_{j+m}x_{j+m}\ a_{i+m}p$ of w^* exists where $I_{last}(p) = 0$ and $|t| = |a_ia_{i+1} \dots a_{i+m}| + 2|x_jx_{j+1} \dots x_{j+m-1}x_{j+m}| - 2 = 2k + 1 + 2(m + 1) - 2 = 2(k + m) + 1$. Consequently, the size of the subword t is odd and $I_{odd}(p)$ is odd. ■

Theorem 4.10. For a given double occurrence word w and a subword $p = a_ib_{i+1} \dots b_{i+m}\ a_j$ of w where $a_i = a = a_j$ for some $i, j \in \mathcal{N}$, the size of the subword $b_{i+1}b_{i+2} \dots b_{i+m}$ is odd if and only if i and j have the same parity.

Proof. Suppose $a \in w$, $a_i = a = a_j$ for some $i, j \in \mathcal{N}$, $j > i$ and the size of the subword $t = b_{i+1}b_{i+2} \dots b_{i+m}$ is odd. Set $p = a_ita_j$. Then $|p| = |a_ita_j| = |t| + 2 = j - i + 1$ is odd and $|p| = 2n + 1$ for some $n \in \mathcal{N}$. This implies $j - i = 2n$ and hence j and i are either both even or both odd, i.e., i and j have the same parity.

Conversely, suppose for any given subword $p = a_ib_{i+1} \dots b_{i+m}a_j$ of a double occurrence word w, i and j have the same parity.

Case 1. i and j are even:
This means that $i = 2y$, $j = 2x$ for some $x, y \in \mathcal{N}$ and $y < x$. Therefore $2y + |b_{i+1} \dots b_{i+m}| + 1 = 2x$ and $|b_{i+1} \dots b_{i+m}| = 2(x - y) - 1$. Hence, the size of the subword $b_{i+1} \dots b_{i+m}$ is odd.

Case 2. i and j are odd:
For some $x, y \in \mathcal{N}$ and $x > y$, $i = 2y+1$, $j = 2x+1$. Then we have $|b_{i+1}b_{i+2} \dots b_{i+m}| = 2x + 1 - (2y + 1) - 1 = 2(x - y) - 1$ which implies the subword is of odd length. ■

Theorem 4.11. Given a double occurrence word w with a symbol $p \in w$ such that $p_i = p = p_j$ in the order and $I_{last}(p)$ is even. Then $I_{odd}(p)$ is odd if and only if i and j have the same parity.

Proof. Suppose $p_i = p = p_j$, $j > i$ and $I_{last}(p)$ is even. There exists a subword $p_i t_{i+1} t_{i+2} \dots t_{i+m} p_j$ of w with $I_{last}(p) = 0$ or 2. Let A be the set of all symbols of $I_{last}(p)$ and $B = \{t_{i+1}, t_{i+2}, \dots, t_{i+m}\} \backslash A$. Given that $I_{odd}(p)$ is odd, we have $|B| = 2n + 1$ for some $n \in \mathcal{N}$.

This implies

$$|t_{i+1} t_{i+2} \dots t_{i+m}| = |B| + |A| = \begin{cases} 2n+3, & \text{if } |A| = 2 \\ 2n+1, & \text{if } |A| = 0 \end{cases}$$

and then the size of the subword $p_i t_{i+1} t_{i+2} \dots t_{i+m} p_j$ equals

$$j - i + 1 = \begin{cases} 2n+5, & \text{if } |A| = 2 \\ 2n+3, & \text{if } |A| = 0 \end{cases}$$

which implies $j - i = 2(n+2)$ or $j - i = 2(n+1)$. Thus j and i have the same parity.

Conversely suppose $p_i = p = p_j$ and $I_{last}(p)$ is even and i and j have the same parity, i.e., i and j are both even or both odd. We show that $I_{odd}(p)$ is odd. Consider the subword $w_1 = p_i t_{i+1} t_{i+2} \dots t_{i+m} p_j$ of a double occurrence word w.

Then the size of w_1, $|w_1| = j - i + 1$ is odd and $|t_{i+1} t_{i+2} \dots t_{i+m}|$ is also odd. Given that $I_{last}(p)$ is even and from the fact that $I_{odd}(p) = |w_1| - I_{last}(p)$, it follows $I_{odd}(p)$ is odd. ■

Theorem 4.12. Given a double occurrence word w with a symbol $p \in w$ such that $p_i = p = p_j$, $j > i$ and $I_{last}(p)$ is odd. Then i and j have opposite parity if and only if $I_{odd}(p)$ is odd.

Definition 4.13. Two distinct symbols in a double occurrence word w are said to be *interlaced* if each appears precisely once between the two occurrences of the other.

Example 4.14. Consider the double occurrence word $w = 12314243$. The symbols 1 and 3 are interlaced but 3 and 4 are not.

Theorem 4.15. If a symbol $a \in w$ is odd then at least one of the symbol $b \in w$ is interlaced with a.

5. Conclusion

In the sections above, we introduced loop-saturated assembly graphs and their properties associated with the assembly numbers, the size of the graphs and the index number. We also showed some sequences of numbers that are not included in The On-Line Encyclopedia of Integer Sequences. Now we conclude the paper with some open questions.

Let $n \geq 1$ be an integer and let w be the double occurrence word of size n and Γ is a simple assembly graph that corresponds to a double occurrence word w. What is the odd index of Γ? For $a \in w$, what is the maximum of $I_{odd}(a)$? Can we characterize double occurrence words that have the same odd index?

References

[1] A. Alhazov, I. Petre, V. Rogojin, 2008, "Solutions to computational problems through gene assembly," *Journal of Natural Computing* 7, pp. 385–401.

[2] A. Angeleska, May 20, 2009, "Combinatoial Models for DNA Re-arrangment in Ciliates," Ph.D. Thesis, University of South Florida, Tampa, USA.

[3] A. Angeleska, N. Jonoska, M. Saito, 2009, "DNA recombinations through assembly graphs," *Discrete Applied Mathematics,* 157, pp. 3020–3037.

[4] A. Ehrenfeucht, T. Harju, David M. Prescott, G. Rozenberg, 2004, *Computation in Living Cells: Gene Assembly in Ciliates*, ISBN-13: 978-3540407959, Springer.

[5] J. Burns, E. Dolzhenko, N. Jonoska, T. Muche, M. Saito, 2013,"Four-regular graphs with rigid vertices associated to DNA recombination," *Discrete Applied Mathematics*, doi:10.1016/j.dam.2013.003

[6] J. Burns, Tilahun M., 2011, "Counting irreducible double-occurrence words," *Proceedings of the Forty-second Southeastern International Conference on Combinatorics, Graph Theory and Computing*, Congr. Number, 207, pp. 181–196.

[7] Lena C. Folwaczny and Louis H. Kauffman, 8 Nov 2012, "A Linking Number Definition of the Affine Index Polynomial and Application," arXiv:1211.1747v1 [math.GT].

[8] Rade Doroslovacki, Olivera Markovic, 2000, "n-words over any alphabet with forbidden any 3-subwords," *Novi Sad J. Math.*, Vol. 30, No. 2, pp. 159–163.

[9] Tilahun Muche, July, 2012, "Hamiltonian Sets of Polygonal Paths in 4-Valent Spatial Graphs," Ph.D. Thesis, University of South Florida, Tampa, USA.

[10] The On-Line Encyclopedia of Integer Sequences, http://oeis.org/.

[11] Masahico Saito, 2010, Chirality of Assembly Graphs, math. usf. edu. saito.