Where graph analysis meets applications

Denis TRYSTRAM

Maths for Computer Science – *Home Work MOSIG 1 – 2017*

DEADLINE nov.24 midnight.
Please, use the following NAMING CONVENTION:
subject: HomeWorkMaths
files (in pdf only): HomeWorkMaths(your lastname).pdf
Send the file using your official email (INP or UGA).

The purpose of this work is to study some properties of a specific class of graphs and to apply the analysis to two different applications, namely a card trick and DNA assembling.

1 Definition and basic properties

Let consider the directed graph DG defined as follows: the k-dimensional graph $DG(k)$ is built over the alphabet $\{0,1\}$. There are 2^k vertices and thus, the vertices are represented uniquely by k symbols. The directed edges (also called arcs) link vertex $(x_0x_1...x_{k-1})$ with 2 vertices by shifting all its symbols by one position to the left and adding a new symbol 0 and 1 at the end of the representation, namely $(x_1...x_{k-1}0)$ and $(x_1...x_{k-1}1)$.

Notice that this graph may have loops.

Draw $DG(2)$ and $DG(3)$

Show that $DG(k)$ is eulerian, $\forall k$ where the definition for directed graphs is natural extension of non-directed graphs.

- $G = (V, E, k)$ is Eulerian iff G is fullly connected and for every vertex $x \in V$, $\delta^-(x) = \delta^+(x)$ (there are as many in-coming as out-coming edges).
- G is Eulerian iff it is the union of arc-disjoint directed cycles

An interesting property is that $DG(k)$ can be expanded to the graph $DG(k+1)$ by a one-to-one correspondence of the arcs into vertices. This transformation of a graph DG is known as the *line digraph*, it will be denoted by LDG in the following. The procedure is as follows.

- The set of vertices of $LDG(k)$ corresponds to the arcs of $DG(k)$.

- The set of arcs in $LDG(k)$ contains all the links between adjacent arcs in $DG(k)$.

Show explicitly this transformation on $DG(2)$.
Show more generally that $LDG(k)$ is $DG(k+1)$.

Show that if $DG(k)$ is eulerian, then $LDG(k)$ is Hamiltonian.

Let us now study a compact coding of such paths in $DG(k)$.

We give as follows the two first orders: cycle 0011 for k=2 and cycle 00011101 for $k = 3$.
Explain this principle on a figure and give the next cycle.
Show why this coding is efficient compared to standard codings (like Gray codes).

2 A card trick

The idea here is to analyze a card trick using graph theory.

We ask someone to take a card in a deck of 32 cards and to replace it at the same place. Looking at the color of the four precedent cards in the deck, we are able to determine the guessed card!

Explain the trick using an adequate coding and the results of the previous section. Of course, the deck should be prepared by a specific procedure...

3 DNA assembling

In this section, we consider another example that uses the previous results.

Let consider DNA sequences composed of a series of nucleotids, written on the alphabet with the 4 letters $\{A, C, G, T\}$.

In bioinformatics, sequence assembly consists in aligning and merging fragments from a longer DNA sequence in order to reconstruct the original sequence. The reason is that it is impossible to read the whole genomes at once. We usually read many small equal-size pieces called *p-mers*. The whole sequence (of size n) corresponds to an eulerian path in a particular directed graph whose vertices are the words of length p and the arcs are the overlaps of $p-1$ symbols. The whole sequence is obtained by reading the symbols one after the other, where very read must be used in exactly one place in the sequence.

$ACGATACGTAC$ is an example of a complete sequence. Take $p = 3$ and build the corresponding graph.
Give the algorithm for reconstructing the whole sequence.