Physical Mappings

Lecture 7

Theorem Every two alternating eulerian circuits in a bicolored graph G can be transformed into each other by a series of order transformations using exchanges and reflections.

Proof. Let X and Y be two of them. Consider the set of alternating eulerian circuits $\mathcal{C}$ from X by all possible transformations. Let $X^* = \langle x_1, x_2, \cdots, x_m \rangle$ be the one in $\mathcal{C}$ having the longest common prefix with $Y = \langle y_1, y_2, \cdots, y_m \rangle$ and $\langle x_1, x_2, \cdots, x_\ell \rangle = \langle y_1, y_2, \cdots, y_\ell \rangle$ for $\ell \le m$. If $\ell = m$, then the proof follows. Otherwise, let $v = x_\ell = y_\ell$ and $(v, \overset{e_1}{\overbrace{x_{\ell+1}}}) \neq (v, y_{\ell+1}) = e_2$. (第一个不一样的边)

Clearly, e_1 and e_2 are of the same color and e_2 succeeds e_1 in X^*. Now, we have two cases to consider.

Case 1. $e_2 = (x_j, v)$ where $x_j = y_{\ell+1}$ and $j > \ell+1$.

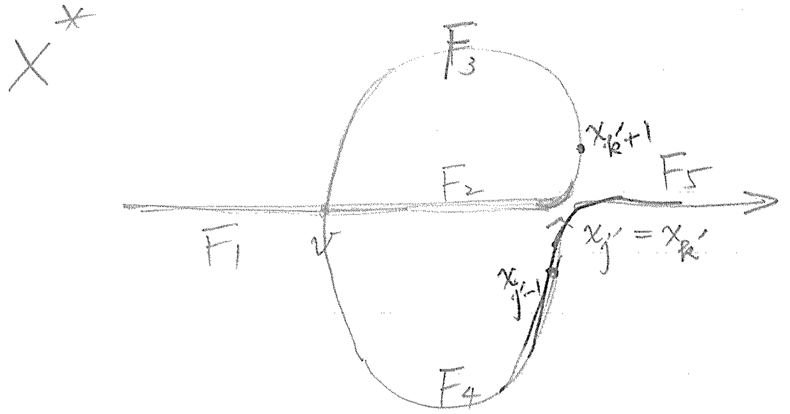
$$X^* = \langle \underbrace{x_1, x_2, \cdots, x_\ell}_{F_1} = v, \underbrace{x_{\ell+1}, \cdots, \overset{y_{\ell+1}}{\overbrace{x_j}}, v}_{F_2}, \underbrace{\cdots, x_m}_{F_3} \rangle$$

This implies that $X^{**} = F_1 \bar{F}_2 F_3$ and Y contains at least the same prefix with $\ell+1$ vertices, a contradiction.

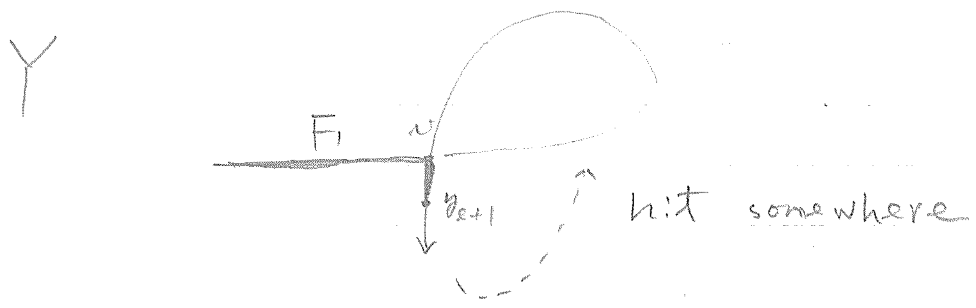
Case 2 $e_2 = (v, x_j)$ in X^* where $x_j = y_{\ell+1}$ and $j > \ell+1$.

$$X^* = \langle \underbrace{x_1, x_2, \cdots, x_\ell}_{X_1}, \underbrace{x_{\ell+1}, \cdots, v}_{X_2}, \underbrace{x_j, \cdots, x_m}_{X_3} \rangle$$

(with $x_\ell = v$ and $x_j = y_{\ell+1}$)



$\overset{\in X_2}{x_{j'}} = \overset{\in X_3}{x_{k'}}$



Consider $(x_{k'}, x_{k'+1})$ and $(x_{j'-1}, x_{j'})$.

Either they use the same color or they are colored with distinct colors.

Case 2.1. $c(x_{k'}, x_{k'+1}) = c(x_{j'-1}, x_{j'})$.

Let $X^{**} = F_1 \bar{F_4} \bar{F_2} \bar{F_3} F_5$.

$F_1 F_2 F_3 F_4 F_5 \longrightarrow F_1 F_2 (F_3 F_4)^- F_5 =$

$F_1 F_2 F_4^- F_3^- F_5 \longrightarrow F_1 (F_2 F_4^-)^- F_3^- F_5$

$= F_1 (F_4^-)^- F_2^- F_3^- F_5 = F_1 F_4 F_2^- F_3^- F_5 .$

Add one more common vertex to prefix, $\rightarrow\leftarrow$.

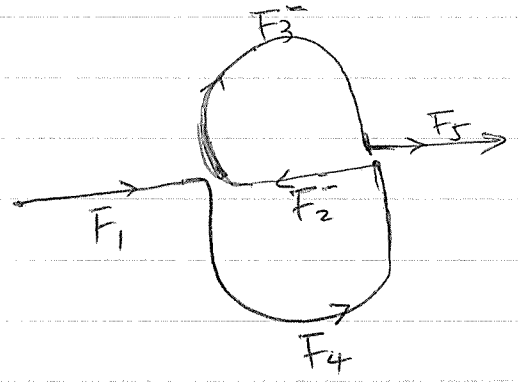
Case 2.2. $c(x_{k'}, x_{k'+1}) \neq c(x_{j'-1}, x_{j'})$.

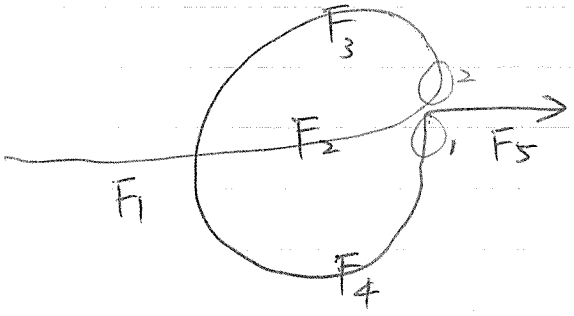
Let $X^{**} = F_1 F_4 F_3 F_2 F_5$ by using order exchange.

Then, we have a contradiction. ▨

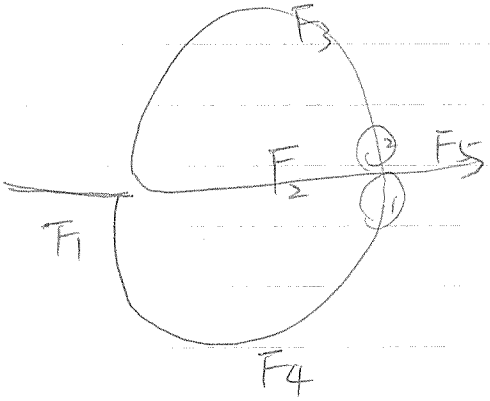
Note

Case 2.1

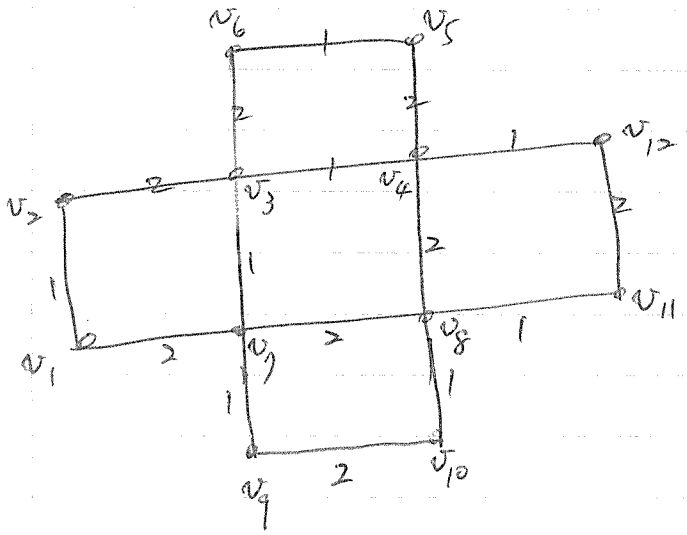




Case 2.2.



An example



$X^* \rightsquigarrow Y$

$X^* = \langle v_1, v_2, v_3, v_7, v_8, \overbrace{v_4, v_5, v_6, v_3, v_4}^{F}, v_{12}, v_{11}, v_8, v_{10}, v_9, v_7 \rangle$

$Y = \langle v_1, v_2, v_3, v_4, v_5, v_6, v_3, v_7, v_8, v_{11}, v_{12}, v_4, v_8, v_{10}, v_9, v_7 \rangle$

$\langle v_1, v_2, \overbrace{v_3, v_7, v_8, v_4, v_3, v_6, v_5, v_4}^{F^-}, v_{12}, v_{11}, v_8, v_{10}, v_9, v_7 \rangle$

$\Downarrow$

$\langle v_1, v_2, v_3, v_4, v_8, v_7, v_3, v_6, v_5, v_4, \cdots \rangle$

$\Downarrow$

$\langle v_1, v_2, v_3, v_4, v_5, v_6, v_3, v_7, \overbrace{v_8, v_4, v_{12}, v_{11}, v_8}, v_{10}, v_9, v_7 \rangle$

$v_8, v_{11}, v_{12}, v_4, v_8, v_{10}, v_9, v_7$ Done!

Physical Maps and Alternating Eulerian Circuits

Consider a physical map given by ordered fragments of single digests A and B, and double digest $C = A + B$: $A = \{A_1, A_2, \cdots, A_n\}$, $B = \{B_1, B_2, \cdots, B_m\}$ and $C = \{C_1, C_2, \cdots, C_\ell\}$. For simplicity, assume that A and B do not cut DNA at the same positions, i.e., $\ell = n + m - 1$.

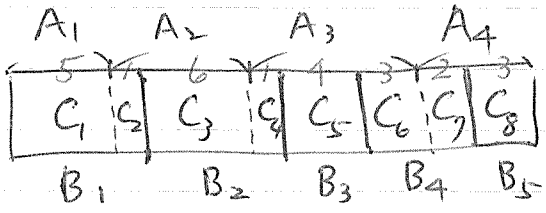


Figure A

Definition (Fork) $(F(B_k))$

A fork of fragment A_i $F(A_i)$, (B_k) is the set of double digest fragments C_j contained in A_i (B_k), i.e.,

$$F(A_i)_{(B_k)} = \{ C_j \mid C_j \text{ is in } A_i \}.$$

e.g. $F(A_2) = \{C_2, C_3\}$, $F(A_3) = \{C_4, C_5, C_6\}$, $F(B_3) = \{C_5\}$, and $F(B_4) = \{C_6, C_7\}$.

(•) A fork containing at least two fragments is called a multifork, e.g. $A_2, A_3, B_4 \cdots$.

(•) Leftmost and rightmost fragments of multiforks are called border fragments.

(•) In example, border fragments are:

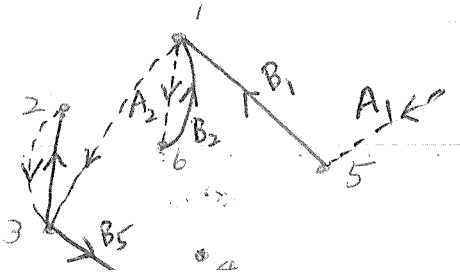
$C_1, C_2, C_3, C_4, C_6, C_7, C_8$, except C_5.

Lemma Every border fragment, excluding C_1 and C_2, belongs to exactly two multiforks $F(A_i)$ and $F(B_j)$. C_1 and C_2 belong to exactly one multifork.

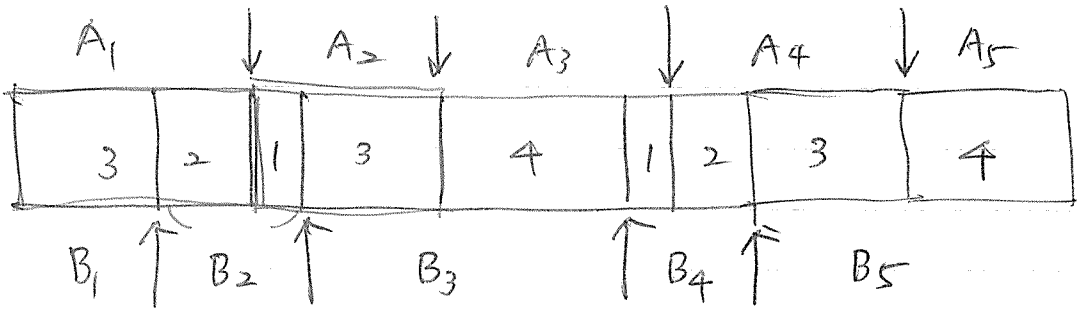
Definition (Fork Graph for double digest $C = A + B$)

Let G_C be the fork graph defined on C. Then $V(G_C)$ is the set of lengths of border fragments and $E(G_C)$ is the set of all multiforks (each multifork is represented by an edge connecting the vertices corresponding to the length of its border fragments).

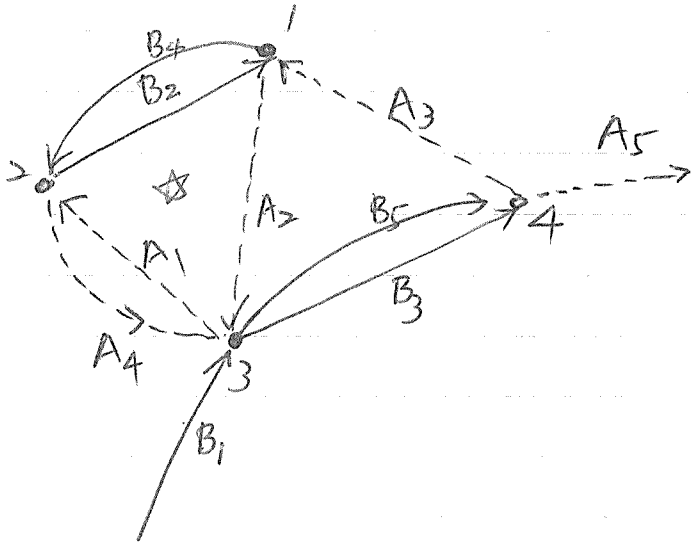
e.g. G_C of Figure A.



Example



Graph



$B_1 (A_1\ B_2\ A_2)\ B_3\ A_3\ B_4\ A_4\ B_5\ A_5$

$B_1\ \bar{A_2}\ \bar{B_2}\ \bar{A_1}\ B_3\ A_3\ B_4\ A_4\ B_5\ A_5$

(Reverse the triangle ☆)

Problems related to eulerian circuits

Known result: If G is a directed eulerian graph, then the number of distinct eulerian circuits, starting at vertex s, is equal to $t^+(G,s) \cdot \prod_{v \in V(G)} (\deg^+(v)-1)!$ where $t^+(G,s)$ is equal to the arborescence converging to s. As a matter of fact $t^+(G,s) = (-1)^{s+t} \det L^+(G)_{s,t}$ where $L^+(G)_{s,t}$ is obtained by deleting row s and column t of $L^+(G) = D^+ - A$. Here,

$$D^+ = \begin{bmatrix} \deg^+(v_1) & & \\ & \deg^+(v_2) & 0 \\ & \ddots & \\ 0 & & \deg^+(v_n) \end{bmatrix} \quad \text{and} \quad A = A(G).$$

D^+: outdegree

$L^+(G) = D^+ - A$: Laplacian of G

Problem 1 Find the number of distinct eulerian circuits in an eulerian graph (not a digraph).

Problem 2 Let G be an eulerian graph. Find the number of distinct (bicolored balanced) labelling.

Problem 3 Let G be a balanced bicolored eulerian graph. Find the number of distinct alternating eulerian circuits starting from a fixed vertex s.

The map assembly problem can be understood in terms of the following analogy. Imagine several copies of a book cut by scissors into thousands of pieces. Each copy is cut in an individual way such that a piece from one copy may overlap a piece from another copy. For each piece and each word from a list of key words, we are told whether the piece contains the key word. Given this data, we wish to determine the pattern of overlaps of the pieces.

Double Digest and Partial Digest techniques allow a biologist to construct restriction (physical) maps of small DNA molecules, such as viral, chloroplast, or mitochondrial DNA. However, these methods do not work (experimentally or computationally) for large DNA molecules. Although the first restriction map of a viral genome was constructed in 1973, it took more than a decade to construct the first physical maps of a bacterial genome by assembling restriction maps of small fragments. To study a large DNA molecule, biologists break it into smaller pieces, map or fingerprint each piece, and then assemble the pieces to determine the map of the entire molecule. This mapping strategy was originally developed by Olson et al., 1986 [257] for yeast and by Coulson et al., 1986 [76] for nematode. However, the first large-scale physical map was constructed by Kohara et al., 1987 [204] for *E. Coli* bacteria.

Mapping usually starts with breaking a DNA molecule into small pieces using restriction enzymes. To study individual pieces, biologists obtain many identical copies of each piece by *cloning* them. Cloning incorporates a fragment of DNA into a *cloning vector*, a small, artificially constructed DNA molecule that originates from a virus or other organism. Cloning vectors with DNA inserts are introduced into a bacterial self-replicating host. The self-replication process then creates an enormous number of copies of the fragment, thus enabling its structure to be investigated. A fragment reproduced in this way is called a *clone*.

For reference

As a result, biologists obtain a *clone library* consisting of thousands of clones (each representing a short DNA fragment) from the same DNA molecule. Clones from the clone library may overlap (overlapping can be achieved by using a few restriction enzymes). After the clone library is constructed biologists want to *order* the clones, i.e., to reconstruct the relative placement of the clones along the DNA molecule. This information is lost in the construction of the clone library, and the process of reconstruction starts with *fingerprinting* the clones. The idea is to describe each clone using an easily determined fingerprint, which can be thought of as a set of "key words" present in a clone. If two clones have substantial overlap, their fingerprints should be similar. If non-overlapping clones are unlikely to have similar fingerprints then fingerprints would allow a biologist to distinguish between overlapping and non-overlapping clones and to reconstruct the order of the clones. The following fingerprints have been used in many mapping projects.

- *Restriction maps.* The restriction map of a clone provides an ordered list of restriction fragments. If two clones have restrictions maps that share several consecutive fragments, they are likely to overlap. With this strategy Kohara et al., 1987 [204] constructed a physical map of the *E. coli* genome.

- *Restriction fragment sizes.* Restriction fragment sizes are obtained by cutting a clone with a restriction enzyme and measuring the sizes of the resulting fragments. This is simpler than constructing a restriction map. Although an unordered list of fragment sizes contains less information than an ordered list, it still provides an adequate fingerprint. This type of fingerprint was used by Olson et al., 1986 [257] in the yeast mapping project.

- *Hybridization data.* In this approach a clone is exposed to a number of probes, and it is determined which of these probes hybridize to the clone. Probes may be short random sequences or practically any previously identified piece of DNA. One particularly useful type of probe is the *Sequence Tag Site* (STS). STSs are extracted from the DNA strand itself, often from the endpoints of clones. Each STS is sufficiently long that it is unlikely to occur a second time on the DNA strand; thus, it identifies a unique site along the DNA strand. Using STS mapping, Chumakov et al., 1992 [68] and Foote et al., 1992 [111] constructed the first physical map of the human genome.