- $c_1 = c_2 = \ldots = c_m = 1$
- $\lambda_0 = \lambda_1 = \ldots = \lambda_m = 0$ and $\Lambda_0 = \Lambda_1 = \ldots = \Lambda_m = \infty$
- $h(t_i, t_j) = \begin{cases} 1 & \text{if } \{v_i, v_j\} \in E \text{ and } S(t_i) \neq S(t_j) \\ 0 & \text{otherwise} \end{cases}$
- b as lower bound for the value of $\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} h(t_i, t_j)$

Lemma 5.40.
Assume that parameters are fixed as above.
(a) From a cut $V = V_1 \cup V_2$ with at least b many edges from E between V_1 and V_2 we obtain a threading $t_1, \ldots, t_m$ with score at least b.
(b) From a threading $t_1, \ldots, t_m$ with score at least b we obtain a cut $V = V_1 \cup V_2$ with at least b many edges from E between V_1 and V_2.
Thus, (V, E) has a cut with at least b many edges between its parts if and only if the constructed instance of PT has a threading with score at least b. This is the desired polynomial reduction from MAX-CUT to PT.

Proof. (a) Consider cut $V = V_1 \cup V_2$ with at least b many edges from E between V_1 and V_2. To thread the i^{th} core segment into $S = (01)^m$ consider node v_i. In case that $v_i \in V_1$ choose bit 0 of the i^{th} substring (01) of S as start position for the i^{th} core segment. In case that $v_i \in V_2$ choose bit 1 of the i^{th} substring (01) of S as start position for the i^{th} core segment. For each edge $\{v_i, v_j\}$ in E with nodes in different parts of the cut, that is with either $v_i \in V_1$ and $v_j \in V_2$, or with $v_i \in V_2$ and $v_j \in V_1$, we obtain a contribution of 1 to the scoring function as different bits are selected for $S(t_i)$ and $S(t_j)$. Thus the threading also has score at least b.
(b) Let a threading $t_1, \ldots, t_m$ with score at least b be given. By definition of the scoring function there must be at least b contributions 1. Contribution 1 occurs only for pairs t_i and t_j with an edge $\{v_i, v_j\} \in E$ and different bits $S(t_i)$ and $S(t_j)$. Defining $V_1 = \{v_i \mid S(t_i) = 0\}$ and $V_2 = \{v_i \mid S(t_i) = 1\}$ thus defines a cut with at least b edges between V_1 and V_2. □

With Lemma 5.41 the proof of Theorem 5.40 is complete. □

5.5.5 Bi-Clustering

Let (V, W, F) be a bipartite graph with node sets V and W, and set of edges $\{v, w\}$ between certain node pairs $v \in V$ and $w \in W$. A bi-clique consists of subsets $A \subseteq V$ and $B \subseteq W$ such that for all $a \in A$ and $b \in B$ there is an edge $\{a, b\}$ in F. Visualizing bipartite graphs and bi-cliques in a rectangular diagram (Fig. 5.21) makes clear that finding a bi-clique with maximum number of edges $|A|\,|B|$ is exactly the formal problem behind finding a maximum size sub-matrix (after permutations of rows and columns) of a microarray data matrix that consists of entries 1 only.

Lemma 5.41.
Given 3SAT-formula

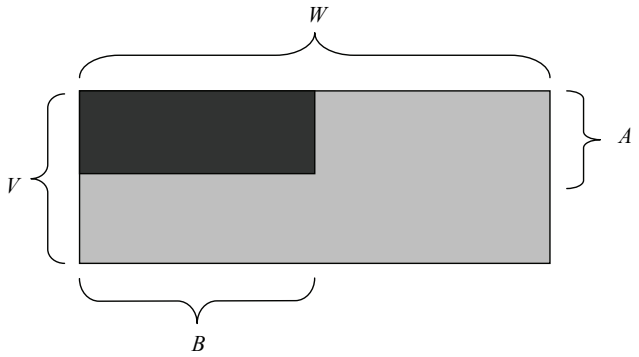

Fig. 5.21. A bi-clique

$$\varphi = (L_{11} \vee L_{12} \vee L_{13}) \wedge (L_{21} \vee L_{22} \vee L_{23}) \wedge \ldots \wedge (L_{n1} \vee L_{n2} \vee L_{n3})$$

with n clauses $(L_{i1} \vee L_{i2} \vee L_{i3})$ each containing three literals L_{i1}, L_{i2}, L_{i3}, an undirected graph (V, E) consisting of $4n$ nodes can be constructed such that φ is satisfiable if and only if (V, E) has a clique of size exactly $2n$ (thus this is a reduction to the variant of CLIQUE where we ask for a clique having exactly half as many nodes as the graph has, called $\frac{1}{2}$-CLIQUE).

$\frac{1}{2}$-CLIQUE
Given undirected graph with $4n$ nodes, does it possess a clique with $2n$ nodes.

Proof. Take a look at the reduction of 3SAT to CLIQUE that was used in Theorem 5.7. There, a graph consisting of $3n$ nodes was used and the clique that occurred had exactly size n. Inserting further n nodes that are completely connected to all of the former $3n$ nodes establishes the assertion of the lemma. □

Theorem 5.42.
$\frac{1}{2}$-CLIQUE is polynomially reducible to BI-CLIQUE, thus BI-CLIQUE is NP-complete, too.

Proof. Given an undirected graph (V, E) with node set $V = \{v_1, \ldots, v_n\}$ of size $n = 2k \geq 16$ with even number k and edge set $E = \{e_1, \ldots, e_m\}$, it can be transformed into a bipartite graph (V, W, F) (first node set of the bipartite graph is indeed the same as the node set from (V, E)) such that (V, E) has a clique C with $|C| \geq k$ if and only if (V, W, F) has a bi-clique consisting of subsets A and B such that $|A|\,|B| \geq k^3 - \frac{3}{2}k^2$ (note that $k \geq 3$ and k is an even number).

We define node set W and edge set F as follows using (somehow unusual, nevertheless admissible) the edges of E as nodes in the second component W together with a number of $\frac{1}{2}k^2 - k$ fresh nodes in W.

$$V = \{v_1, \ldots, v_n\}$$
$$W = \left\{e_1, \ldots, e_m, f_1, \ldots, f_{\frac{1}{2}k^2-k}\right\}$$
$$F = \{\{v_i, e_j\} \mid 1 \leq i \leq n, 1 \leq j \leq m, v_i \notin e_j\}$$
$$\cup \left\{\{v_i, f_j\} \mid 1 \leq i \leq n, 1 \leq j \leq \frac{1}{2}k^2 - k\right\}$$

We show that the desired reduction property holds. In one direction, assume that (V, E) has a clique of size at least k. Take a clique C of size exactly k. Define bi-clique A, B as follows:

$$A = V - C$$
$$B = \left\{f_1, \ldots, f_{\frac{1}{2}k^2-k}\right\} \cup \{e_j \mid 1 \leq j \leq m, e_j \subseteq C\} \ .$$

Thus, B contains all fresh nodes as well as all edges from E that connect nodes of clique C. As C is a clique, there is an edge $\{a, b\}$ in E for any two different nodes a, b in C. Size of the defined bi-clique is thus calculated as follows:

$$|A| = k = |C|$$
$$|B| = \frac{1}{2}k^2 - k + \frac{1}{2}k(k-1)$$
$$|A|\,|B| = k\left(\frac{1}{2}k^2 - k + \frac{1}{2}k(k-1)\right) = k^3 - \frac{3}{2}k^2 \ .$$

In the converse direction, assume that (V, W, F) has a bi-clique consisting of subsets $A \subseteq V$ and $B \subseteq W$ such that $|A|\,|B| \geq k^3 - \frac{3}{2}k^2$ holds. We may assume that B contains all of the fresh nodes (otherwise simply put the missing nodes into B, obtaining again a bi-clique of even larger size) as well as b many of the edges from E as nodes. By definition of edge set F we know that edges e_j occurring in B do not contain any nodes from A. Thus they consist of nodes from $V - A$ only. By renumbering edges in E we may assume without loss of generality that B looks as follows:

$$B = \left\{f_1, \ldots, f_{\frac{1}{2}k^2-k}\right\} \cup \{e_j \mid 1 \leq j \leq b\} \text{ with } e_j \subseteq (V - A) \times (V - A) \ .$$

Now consider node set $C = V - A$. Define $a = |A|$. Thus, the considered bi-clique is located as shown in Fig. 5.22. As C consists of $2k - a$ many nodes, the number of edges connecting nodes from C, and thus also number b can be bounded as follows:

$$b \leq \frac{1}{2}(2k - a)(2k - a - 1) \ .$$

We show that $a \leq k$ follows from this. Assume that $a > k$ holds. It is convenient to further compute with $x = a - k$. By definition of x and from $a \leq 2k$ we conclude that $0 < x \leq k$ holds. Furthermore, $2k - a = k - x$ holds. We obtain a contradiction as follows (the last estimation uses $x \leq k$):