Minimizing Total Weight of Feedback Arcs in Flow Graphs

Marija Mihova, Ilinka Ivanoska, Kristijan Trifunovski, Kire Trivodaliev

Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering,
1000 Skopje, Macedonia
{marija.mihova,ilinka.ivanoska,kire.trivodaliev}@finki.ukim.mk,
kichosh@yahoo.com

Abstract. The problem of determining the minimal weight feedback edge set in an oriented weighted graph when the graph represents a network flow is the focus of this paper. The proposed solution is especially interesting since it can be directly applied in solving the linearization of genome sequence graphs (GSG) which represent a set of individual haploid reference genomes as paths in a single graph and can find frequent human genetic variations, including translocations, inversions, deletions, and insertions. The linearization is essential for the efficiency of storing and traversal as well as visualization of the graph. In this paper properties of the flow function are used to obtain some features of the optimal ordering of nodes when the weight of the feedback edge set is minimal.

Keywords: Two-terminal flow network · Feedback arcs · Linearization · Flow graph.

1 Introduction

In graph theory, a feedback arc set, or feedback edge set in an unweighted directed graph is a set of edges which, when removed from the graph, leave a directed acyclic graph, or DAG. In fact, it is a set containing at least one edge of every cycle in the graph. Finding a minimal set with this property is a key step in layered graph drawing [1]. In weighted graph this operation can be defined as a minimization of the total weight of feedback arcs, which is the minimum weight of the edges such that when removed from the graph, a DAG is obtained. This problem has various applications, most prominent ones being tournaments graphs [5] and genome sequence graph [3]. The decision version of the problem is NP-complete and one of the Richard M. Karp's NP-complete problems [6]. There are some algorithms whose running time is a fixed polynomial in the size of the input graph, but exponential in the number of edges in the feedback arc set [2] or dimension of a cycle space [4], although some special cases of tournament graphs can be solved in polynomial time.

The focus of application in this paper is finding the minimal feedback edge set of a Genome Sequence Graph (GSG) [9], a graph used to represent the human genetic variation into the reference human genome. Each node of a GSG represents a single

DNA base that occurs at an orthologous locus in one or more of the haploid genomes represented. Each arc corresponds to an adjacency that occurs between consecutive instances of bases in the represented genomes, and it is directed according to the default strand direction of the DNA sequence. The weight to each arc is added in order to emphasize the importance of an arc in typical applications running on the graph, for example number of times that the arc is traversed in the reference genomes used to build the graph. The representation of GSG has importance for the efficiency and the effectiveness of the operations such as access, traversal and visualization, all of which depend on the GSG structure/representation, and the best GSGs are ones that have their nodes ordered in a straight line, a structure obtained in a process known as linearization of the graph. Usually the linearization of a sequence graph aims to lessen the total weight of all feedback arcs, i.e. the weighted feedback, as much as possible. On the other hand, a GSG graph can be easily modified into a graph that represent a flow, which means that the inflow into each node representing a genome is equal to the out flow out of any node representing a genome. This characteristic gives specific features to the minimal weighted feedback edge ordering, so pointing and distinguishing some of them is the main interest in this paper.

The rest of this paper is organized as follows. The second section presents the terminology and definitions from network flow theory used in the research. The next section defines the linearization of Genome Sequence Graph as a network flow problem. Section 4 presents our findings in terms of specific features the optimal ordering of nodes has when the weight of the feedback edge set is minimal. The paper is concluded in the fifth section.

2 Flow Function and Minimal Path Vectors

A *flow* in an undirected network $G(V, E, c)$ is a function $f: V \times V \rightarrow R^{+} \cup \{0\}$ that satisfies the following three constraints:

1. *Capacity constraint*: $0 \leq f(u, v) \leq c(\{u, v\})$, for all $u, v \in V$, i.e., the flow of an edge cannot exceed its capacity.

2. *Flow conservation*: for each $v \in V$,

$$f(V, v) - f(v, V) = \sum_{u \in V} f(u, v) - \sum_{w \in V} f(v, w) = \begin{cases} 0, & v \notin \{s, t\} \\ |f|, & v = s \\ -|f|, & v = t \end{cases}, \quad (1)$$

where $|f|$ is the value of the flow. In other words, the total flow entering a node v, $f(V, v)$, must be equal to the total flow leaving node v, $f(v, V)$, $\forall v \in V \setminus \{s, t\}$; the flow leaving s and the flow entering t is equal to the value of the flow.

3. For all vertices $u, v \in V$, if $f(u, v) > 0$, then $f(v, u) = 0$. In other words, each flow uses a given edge only in one direction. It is assumed that if there is no edge $\{u, v\}$, i.e. $\{u, v\} \notin E$, then $f(u, v) = 0$.

In the rest of the paper we will use the term flow graph for a weighted directed graph with flow as a weight function.

Here we give some additional definitions that help define the problem and explain the algorithm.

Definition 1. *Let G(V, E, c) be a two-terminal flow network. For a flow f, we define flow vector* $\overrightarrow{x_f}$ *induced by f by*

$$x_i = f(u, v) = f(e_i), \tag{2}$$

where $e_i = \{u, v\}$.

Definition 2. *Given a two-terminal undirected flow network G(V, E, c), the vector* $\vec{x}$ *is a path vector to level d, d-P, if and only if there is a flow f with a value |f| = d, such that for all edges* $e_i = \{u, v\}$, $f(u, v) \leq x_i$. *In other words,* $\vec{x}$ *is a path vector to level d if and only if a flow d may be delivered in the two-terminal network G(V, E,* $\vec{x}$*).*

Definition 3. *The state vector* $\vec{x}$ *is a minimal path vector to level d, d-MP, if and only if the two-terminal flow network G(V, E,* $\vec{x}$*) has a maximum flow d, and for each* $\vec{y} \leq \vec{x}$, *the two-terminal flow network G(V, E,* $\vec{y}$*) has a maximum flow less or equal to d.*

Theorem 1. *The state vector* $\vec{x}$ *is a d-MP for the two-terminal flow network G(V, E, c) iff* $\vec{x}$ *corresponds to a flow function with flow d and the corresponding graph G(V, E,* $\vec{x}$*) has no cycles [7, 8].*

Corollary 1. *Each 0-MP is a sum of simple cycles* $\overrightarrow{c_k}$

Corollary 2. *If* $\vec{x}$ *is a flow vector for flow d and* $\overrightarrow{c_k}$*are simple cycles then*

$$\vec{y} = \vec{x} - \sum_k \overrightarrow{c_k} \tag{3}$$

is a flow vector. Moreover, $G\left(V, E^{\downarrow \vec{y}}\right)$ *is acyclic iff* $\vec{y}$ *is a d-MP.*

Example 1. Let us consider the graph given on Fig. 1 a). This graph contains cycles, and one of those cycles is <a, c, d, a>. The flow through that cycle is equal to 2. We can take out that flow from the graph, thus obtain the graph presented on Fig. 2 b). The obtained graph contains a cycle again, <b, c, b>, and by removing this cycle we will obtain the graph given on Fig. 2.

If we order the nodes as <s, d, a, b, c, t>, the graph on Fig 2 has not forward edges. Moreover, if we order the nodes of the graph on Fig 1 a), the total weight of the feedback edges will be 3, which is the minimal possible weighted feedback of this graph.

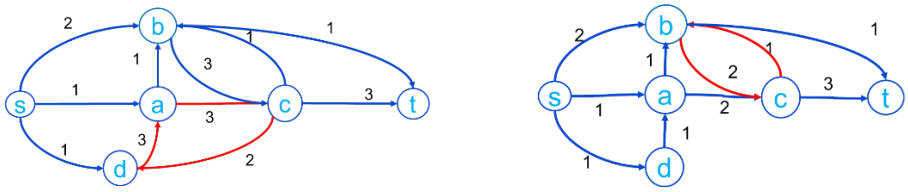

Fig. 1. The graph from Example 1. a) <a, c, d,c> is a cycle with capacity 2, that should be removed from the graph. b) The graph obtained from a).

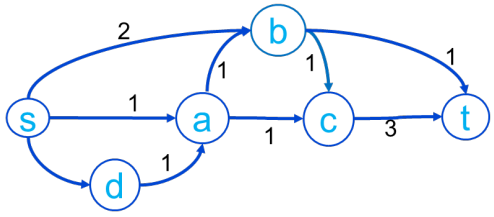

Fig. 2. Acyclic graph obtained by removing cycles from the graph shown on Fig 1.

3 Linearization of Genome Sequence Graphs

Genome Sequence Graph (GSG), Fig. 3, is a representation of DNA sequences that helps to incorporate human genetic variation into the reference human genome (translocations, inversions, deletions, and insertions.) [9].

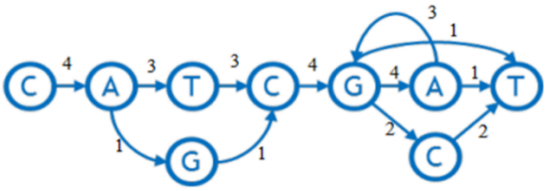

Fig. 3. Genome sequence graph for CATCGCT, CATCGT, CAGCGAT and CATCGAGAGAGCT.

One of the most important operations in GSGs is graph linearization, essential both for efficiency of storage and access, as well as for natural graph visualization and compatibility with other tools. The linearization of a sequence graph aims to lessen the total weight of all feedback arcs, i.e. the weighted feedback, as much as possible. In the rest of the paper we will denote linear ordered set of nodes by $<v_1,\ldots, v_n>$.

The GSG graph can be modified to a flow function in network in the following manner. First, we add special (artificial) source and a special (artificial) sink to the set of nodes. For each node that represents a starting DNA base, we add an arc from the

source to it, with a weight equal to the number of genomes that start with it. For each node that represents an ending DNA base, we add a weighted arc from that node to the sink. The weight of that arc is equal to the number of genomes that end with that node, Fig. 4.

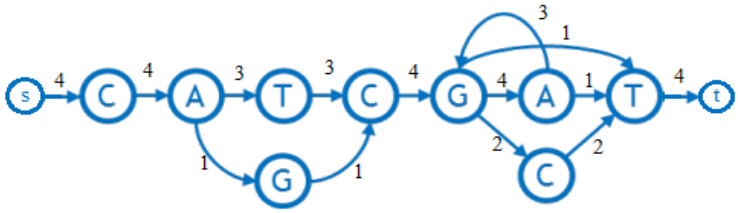

Fig. 4. Modification of the initial DNA graph by adding source and sink.

This way we obtain a graph whose weight function is in fact a flow function, because the total flow out of the source is equal to the total flow entering the sink, and for all other nodes, total flow leaving a node is equal to the total flow entering a node [7, 8]. In fact, such a graph represents a network flow function with flow equal to the number of genome sequences. The max-flow, M, of this network is equal to the total flow out of the source, so any minimal path vector for level M, i.e. minimal flow function for flow M, is an undirected graph. The features of these minimal paths can help in minimizing the weighted feedback of the graph.

4 Total Weight Feedback Arcs and min-paths

The first main new theorem in this part says that if some ordering of the nodes corresponds to the minimal weight of feedback edges, then this ordering also corresponds to a topological sort of some minimal path.

Theorem 2. *Let $G(V, E, c)$ be a flow graph with flow d, where $|V| = n$, and let $<s=v_1, v_2, \ldots, t=v_n>$ be an ordering of nodes with a minimal weight of feedback edges. Then there is a d-MP $\vec{x}$ for which $<v_1, \ldots, v_n>$ is a topological sort of the graph $G(V, E, \vec{x})$.*

Proof. Assume that there is an ordering of nodes with a minimal weight of feedback edges such that the flow in the graph after removing forward edges is $d' < d$. In the graph obtained from G by removing feedback edges, we can find some d'-MP $\vec{x}$. It is obvious that the graph $G'(V, E, c\text{-}x)$ is a flow graph of level $d - d'$ and that any flow of level $d - d'$ uses a collection of feedback edges (v_{j1}, v_{i1}), (v_{j2}, v_{i2}), …, (v_{jk}, v_{ik}), is a collection of feedback edges with weights $d_1, d_2, \ldots, d_{1k}$. Moreover, if we create a new graph by adding a source s' and a sink t' such that $G''(V'=\{\{\min\{v_{im}\},\ldots, \max\{v_{jm}\}\}, E_{/V\cup\{s',t'\}}, c_{/V\cup\{s',t'\}},)$, Where $c_{/V\cup\{s',t'\}}$ is: $\{c(u_k,v_l)|\ k<l\}\cup\{c(s',\ v_{jl})=d_l|\ (v_{jl},\ v_{il})$ is a feedback edge$\}\cup\{c(v_{il},\ t')=d_l|\ (v_{jl},\ v_{il})$ is a feedback edge$\}$, then the flow from s' to t' is less than $d_1+d_2+\ldots+d_{1k}$, which is in contradiction with Lemma 2 given below. □

Lemma 1. *Let G(V, E, c) be a flow graph where |V| = n, and let <s=v_1, v_2,..., t=v_n> be nodes ordering with a minimal weight of feedback edges. If (v_j, v_i) is a feedback edge with weight d_1, there is a flow d_1 from v_i to v_j in the subgraph $G_1(V_1=\{v_i,\ldots, v_j\}, E_{/V1}, \vec{x}_{/V1})$.*

Proof. Suppose that the nodes ordered such that the weight of feedback edges is minimal, but we have a feedback edge (v_j, v_i) with weight d_1, and there is not a flow d_1 from v_i to v_j in the subgraph $G_1(V_1=\{v_i,\ldots, v_j\}, E_{/V1}, \vec{x}_{/V1})$. Then there is a cut that decomposes V_1 into sets A and B such that $v_i \in A$, $v_j \in B$, and $\sum_{u_k \in A, u_t \in B} |x_{kt}| < d_1$($|x_{kt}|$ is the weight of the edge from u_k to u_t). If we swap the nodes of the set A and set B, retaining the order of the nodes inside each of them the same as before, the situation with the feedback edges become the following:

- All edges inside the sets $\{v_1,\ldots, v_{i-1}\}$, $\{v_{j+1},\ldots, v_n\}$, A and B remain in the same orientation, so we do not have any changes in weight of the feedback edges between these edges.
- All edges between $\{v_1,\ldots, v_{i-1}\}$ and A, $\{v_1,\ldots, v_{i-1}\}$ and B, $\{v_{j+1},\ldots, v_n\}$ and A and $\{v_{j+1},\ldots, v_n\}$ and B remain in the same orientation, so again we do not have any changes in weight of the feedback edges between these edges.
- The between from A and B are turning in opposite direction, so because $\sum_{u_k \in A, u_t \in B} |x_{kt}| < d_1$, we will obtain feedback edges with weight $\sum_{u_k \in A, u_t \in B} |x_{kt}|$ which is lower than the weight d_1 of the feedback edge we have before this transformation.

Hence, we proved that we can obtain a graph with a smaller total weight of the feedback edges, which is in contradiction with our assumption that the starting ordering have minimal weight of feedback edges.

Let us illustrate the proof of this theorem by following example:

Example 2. Let us consider the graph given on Fig 5. One random ordering of the nodes is given on the picture. The edge (d, a) is one feedback edge that lies on the cycle <d, a, b, c, d> with capacity 1. Removing this edge, together with the nodes *s* and *t*, the residual graph has min cut (A={a, b, c}, B={d}). By swapping the nodes from the sets A and B, retaining the order of the nodes inside each of them the same as before, we will obtain the ordering illustrated on Fig 6. It is obvious that the new order has less weight of the feedback edges.

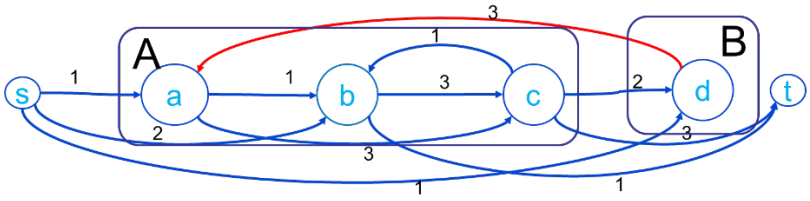

Fig. 5. The graph from Example 2. In this ordering the edge (d, a) is a feedback edge making the cycle <d, a, b, c, d> with capacity 1.

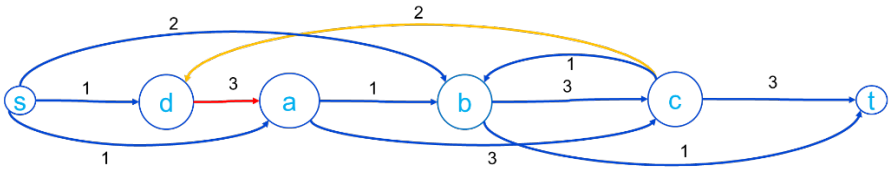

Fig. 6. New linearization of the graph from Fig 5, obtained by swapping the sets A and B.

The new linearization is the best possible, since we cannot further reduce the weight of the feedback arcs. In fact, one feedback edge is (c, d), but the flow in the graph G'= ({d, a, b, c}, $V/_{\{d, a, b, c\}}$) with source d and sink c is greater than 2. We have the same situation with the feedback edge (c, b), since the flow in the graph G''= ({b, c}, $V/_{\{d, a, b, c\}}$) with source b and sink c is greater than 1. Because we can remove these cycles from the graph, Fig 7., we may conclude that further improvement is not possible.

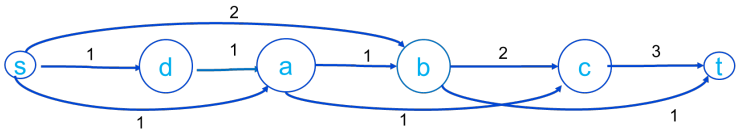

Fig. 7. The graph obtained from the one from Fig 6 by removing all cycles.

The following lemma is generalization of the Lemma 1, and its proof is very similar with the proof of Lemma 1, so we will leave it without proof, but we will illustrate it by an example.

Lemma 2. *Let G(V, E, c) be GSG and let* $<v_1,\ldots, v_n>$ *be nodes ordering with a minimal weight of feedback edges. If* (v_{j1}, v_{i1}), (v_{j2}, v_{i2}), …, (v_{jk}, v_{ik}), *is a collection of feedback edges with weights* $d_1, d_2,\ldots, d_{1k}$, *there is a flow* $d_1+d_2+\ldots+d_{1k}$, *from s' to t' in the graph* $G'(V'=\{\{min\{v_{im}\},\ldots, max\{v_{jm}\}\}, E/_{V'\cup\{s',t'\}}, c/_{V'\cup\{s',t'\}})$, *Where* $c/_{V'\cup\{s',t'\}}$ *is:* $\{c(u_k,v_l)|\ k<l\}\cup\{c(s', v_{jl})=d_l|\ (v_{jl}, v_{il})$ *is a feedback edge*$\}\cup\{c(v_{il}, t')=d_l|\ (v_{jl}, v_{il})$ *is a feedback edge*$\}$.

Example 3. The ordering of the graph on Fig 8 is such that the total weight of feedback edges is 4. We modify it such that we delete the forward edges, add source *s*' and sink *t*', and add edges (*s*', *a*) and (*c, t*') with weight 3, and (*s*', *b*) and (*d, t*') with weight 3. The flow through the second graph is 3, i.e. smaller than the total weight of feedback edges. The cut is between nodes *b* and *c*, so we will swap sets {*a, b*} and {*c, d*}. The obtained order has minimal weight of the feedback edges, equal to 3.

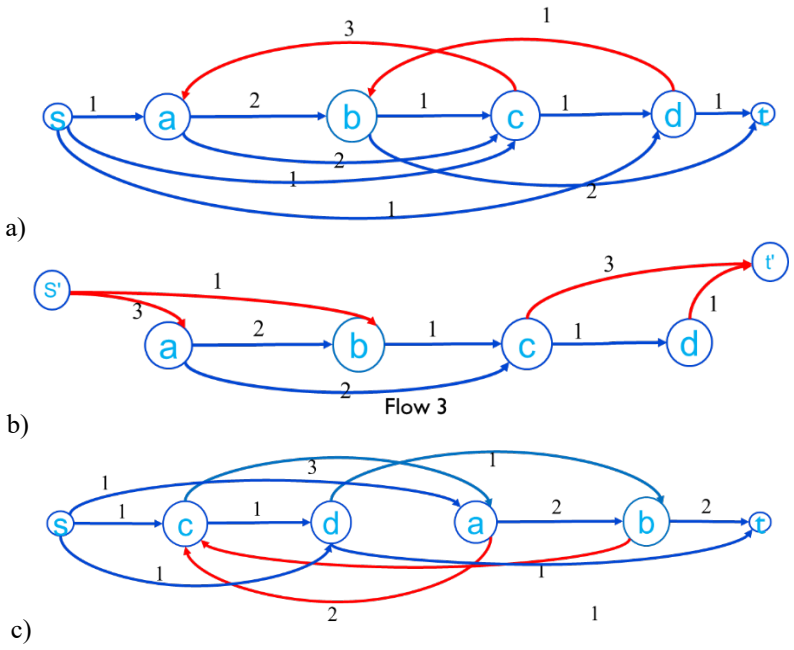

Fig. 8. The graph with weight 4 of the feedback edges and the subgraph induced from the forward edges, modified according to Lemma 2. The last figure illustrates the same graph after swapping sets {a, b} and {c, d}, leading to a graph with feedback edges weight equal to 3.

Note that the Lemma 2 does not provide sufficient conditions for minimal weight feedback edge, i.e. it is not proven that the ordering in which all collection of feedback edges satisfied the condition of Lemma 2 is a minimal weight feedback edge, which is illustrated by Example 4.

Example 4. On the ordering on fig 9 a) we have 2 feedback edges, which means 3 possible collections: {(d, a)}, {(f, c)}, {(d, a), (f, c)}. The conclusions of the Lemmas 1 and 2 are satisfied, but we have more optimal ordering illustrated on Fig 9 b). In fact, following the approach of the Lemma 2, the graph G' induced from the graph on Fig 9 a) is given on Fig 9 c). The characteristic property of these ordering is that in the graph G' defined in Lemma 2 we can find flow of level 2, but the edges used in that flow, together with the feedback edges, form only one simple cycle instead 2.

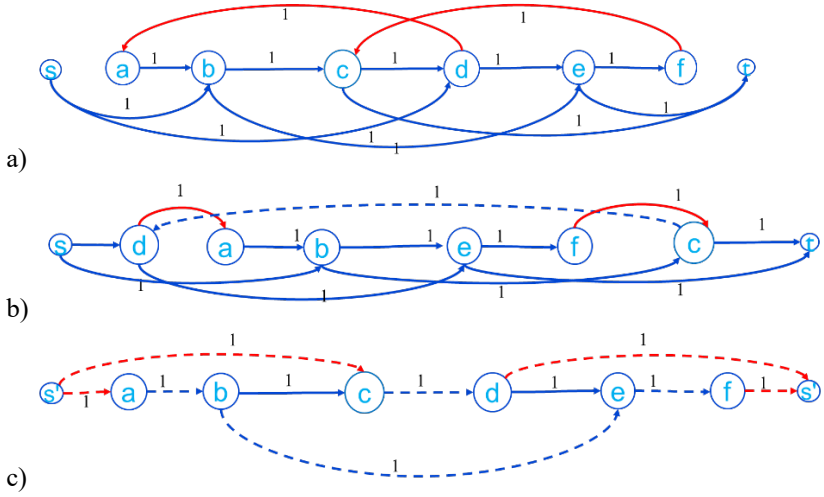

Fig. 9. a) The ordering of a graph with total weight 2 of the feedback edges. b) Ordering of the same graph a) with total weight 2 of the feedback edges. c) The graph G' defined in Lemma 2.

The last example leads to the idea of the following theorem, which gives sufficient condition whether given ordering can lead to minimal weight feedback edges set:

Theorem 3. *Let G(V, E, c) be GSG and let* $<v_1,\ldots, v_n>$ *be nodes ordering such that* $(v_{j1}, v_{i1}), (v_{j2}, v_{i2}), \ldots, (v_{jk}, v_{ik})$ *are all feedback edges having weights* $d_1, d_2,\ldots, d_{1k}$. *We form a flow network with source s' and sink t' as* $G'(V'=\{\{min\{v_{im}\},\ldots, max\{v_{jm}\}\}, E_{/V'\cup\{s',t'\}}, c_{/V'\cup\{s',t'\},})$, *where* $c_{/V'\cup\{s',t'\}}$ *is:* $\{c(u_k,v_l)|\ k<l\}\cup\{c(s', v_{jl})=d_l|\ (v_{jl}, v_{il})$ *is a feedback edge*$\}\cup\{c(v_{il}, t')=d_l|\ (v_{jl}, v_{il})$ *is a feedback edge*$\}$. *Moreover assume that there is a flow* $d_1+d_2+\ldots+d_{1k}$ *in G', and the graph induced from that flow function and the feedback edges can be present as a collection of* $d_1+d_2+\ldots+d_{1k}$ *simple cycles. Then that ordering is an ordering with a minimal weight of feedback edges.*

Proof. If the conditions of the Theorem follows, then any feedback edge is part of exactly one cycle, and for each simple cycle there is exactly one feedback edge. This means that it is impossible to remove less feedback edges from the graph and still obtain a directed acyclic graph as a result.

The proofs of the Lemmas given in this section leads to the following algorithm for reducing total feedback weight:

Algorithm 1.
Step 1. Order the nodes of the graph G in topological sort corresponding to some *d*-MP, where *d* is a maximum flow of the network.
Step 2. While there is a collection of feedback edges that do not satisfied the conditions in Lemma 2
Step 2.1. Create a graph *G'* given by Lemma 2 and find the min cut (*U*, *V*).

Step 2.2. Swap the sets of nodes U and V to obtain new ordering of the nodes in the graph G.

It is obvious that this algorithm is much better than its competing algorithm proposed in [3] in terms of complexity, while we guarantee that the total weight of the feedback arcs will be at most as the one obtained in [3]. Even in the examples used in [3] to illustrate their procedure, the algorithm proposed in this research finds the optimal ordering either after step 1 or in only one iteration in step 2, which is significantly less than what the procedure of [3] requires.

5 Conclusion and Future Work

This paper presents some theoretical results that show the connection between the ordering of nodes of a directed graph with minimal feedback weight and the flow function, for their corresponding weighted flow graphs. The necessary conditions for a node ordering to be optimal are given, and furthermore proofs of Lemmas 1 and 2 provide a procedure for reducing the weight of the feedback edges. The results lead to an algorithm for reducing feedback edges and obtaining a linearized graph. Theorem 3 gives the sufficient condition for an optimal ordering, however the procedure for obtaining one cannot be drawn in a straightforward manner. The next steps in furthering this research will include defining a procedure that will allow for determining the optimal solution taking into account the sufficient conditions theorem, as well as constructing a more detailed and efficient algorithm for reducing the number of feedback edges by selectively choosing which starting feedback edge subsets will be taken into consideration. The final steps will include complexity analysis of the new optimal algorithm and experimental verification for its practical usability.

References

1. Bastert, O., Matuszewski, C.: Layered drawings of digraphs. In: Drawing graphs, pp. 87–120. Springer (2001)
2. Chen, J., Liu, Y., Lu, S., O'sullivan, B., Razgon, I.: A fixed-parameter algorithm for the directed feedback vertex set problem. Journal of the ACM (JACM) **55**(5), 21 (2008)
3. Haussler, D., Smuga-Otto, M., Eizenga, J.M., Paten, B., Novak, A.M., Nikitin, S.,Zueva, M., Miagkov, D.: A flow procedure for linearization of genome sequence graphs. Journal of Computational Biology 25(7), 664–676 (2018)
4. Hecht, M.: Exact localisations of feedback sets. Theory of Computing Systems **62**(5),1048–1084 (2018)
5. Karpinski, M., Schudy, W.: Faster algorithms for feedback arc set tournament, Kemeny rank aggregation and betweenness tournament. In: International Symposium on Algorithms and Computation. pp. 3–14. Springer (2010)
6. Karp, R.M.: Reducibility among combinatorial problems. In: Complexity of computer computations, pp. 85–103. Springer (1972)

7. Mihova, M., Stojkovikj, N., Jovanov, M., Stankov, E.: On maximal level minimal path vectors of a two-terminal network. International Journal of Olympiads in Informatics, **8**(1), 133 – 144 (2014)
8. Mihova, M., Stojkovikj, N., Jovanov, M., Stankov, E.: Maximal level minimal pathvectors of a two-terminal undirected network. IEEE Transactions on Reliability **65**(1), 282–290 (2015)
9. Paten, B., Novak, A. and Haussler, D., 2014. Mapping to a reference genome structure. arXiv preprint arXiv:1404.5010 (2014)