J Autom Reasoning (2015) 55:295–306
DOI 10.1007/s10817-015-9337-1

Improving Legibility of Formal Proofs Based on the Close Reference Principle is NP-Hard

Karol Pąk[1]

Received: 15 September 2013 / Accepted: 23 June 2015 / Published online: 12 July 2015

Abstract Proof development in proof assistants such as HOL, Coq, Mizar, etc. is an activity where authors usually produce proofs by typing out proof scripts or system tactics. Quite frequently, however, authors also have to read existing proof scripts, either to imitate smart proof pieces, or to refactor fragments of reasoning to make some theorem stronger, more easily applicable and so on. Therefore, it is important to develop techniques to improve legibility of proofs, since it directly affects productivity of script writers. To analyze the legibility of natural deduction proofs, we investigate proof graphs that represent the flow of information in given reasoning. Our analysis of the information flow leads to methods of improving proof readability based on Behaghel's First Law, which states that in legible text relevant pieces of information must occur close to each other. The presented method maximizes the number of close connections between premises and steps that use these steps as justification. In this paper we show that our optimization method is NP-hard.

Keywords Natural deduction · Legibility · NP-completeness

1 Introduction

Analyzing declarative natural deduction proofs developed with proof assistants, one may conclude that their legibility often seems to be of secondary importance to their authors. Computer assisted proof development frameworks can check of the proofs scripts created in this way, according to the opinion of some proof writers, is extremely difficult or

The paper has been financed by the resources of the Polish National Science Centre granted by decision no DEC-2012/07/N/ST6/02147

✉ Karol Pąk
pakkarol@uwb.edu.pl

[1] Institute of Informatics, University of Białystok, Białystok, Poland

even impossible. Still, the experience of big proof development efforts shows that adapting existing proofs is unavoidable and requires reading proof scripts [7].

There are proof scripts whose authors spend a lot of time over their readability [2, 11–14]. However, an analysis of proof scripts, especially long and more complex ones, leads to a conclusion that the readability of proof scripts in general might be very far from the acceptable level of readability. This concerns especially systems such as Isabelle/Isar [28] or Mizar [8], where the proof script language is close to the natural language [26].

Clearly, authors of formal proof scripts can manually try to improve readability of their works in a similar way some authors of informal mathematical proofs make them more readable than others. [9]. However, the digital form of structured formal proofs enables not only automatic correctness verification, but also automatic enhancement of proof scripts. Systems such as Mizar are being developed in many directions to improve proof scripts collected in Mizar Mathematical Library (MML) [20]. In particular,

- (i) the visualization of proof scripts in HTML format is being improved [27],
- (ii) the new Mizar language constructions [16, 17, 19] that stem from informal deductions are implemented into new versions of MML [10], respecting the license requirements [1],
- (iii) there are experiments with strengthening the Mizar inference checker by implementing selected computer algebra capabilities in order to reduce user input and shorten MML texts [21],
- (iv) methods of rebuilding the reasoning structure to extract lemmas have been developed [23].

The aim of this article is to consider methods that improve the legibility of natural deduction proofs by changing the order of reasoning steps. Results obtained by initial experiments with step order manipulation were implemented in MML version `4.127.1060` [22]. These results were positively received by Mizar users despite the fact that the experiment used a simple greedy algorithm. Therefore, it seems important to further examine methods that can be used for this goal.

Methods that reorganize the order of steps focus mainly on the location of the information used to justify a given step. Clearly, premises that are used to justify a step have to be previously derived in the proof, but this information can be located somewhere far away in the proof, or within a close neighborhood of the step that refers to it. We focus mainly on this aspect of legibility. According to models of cognitive perception of read material and the Behaghel's First Law, we follow this principle: *elements that belong close together intellectually will also be placed close together* [3]. With Behaghel's law in mind, we assume that a reference which connects a premise and a step that uses it in the justification is more comprehensive if between the premise and the step only few other steps occur. The exact number of these other intermediary steps that can appear without significant loss of readability is different for different readers. But the opinions of several users of the Mizar database seem to suggest that the number might be fixed for every person.

In this paper we study the computational complexity of methods that have origins in Behaghel's law. In Section 2, we formulate various problems of improving the legibility of natural deduction proofs in terms of DAGs. Then, in Section 3, we show that formulated problems are NP-complete. In Section 4 we reformulate one of these problems in terms of Hamiltonian paths. Finally, Section 5 concludes the paper and discusses the future work.

2 Formulation of Behaghel's Law Determinants

To formulate legibility criteria we first need to set up the terminology and notation. Let $G = \langle V, E\rangle$ be a DAG and a vertex $u \in V$. We assume that G is without self-loops i.e. without edges that connect a vertex to itself. We use the following notation:

$$\begin{aligned} N_G^-(u) &:= \{v \in V : \langle v, u\rangle \in E\} \qquad \text{(incoming arcs)},\\ N_G^+(u) &:= \{v \in V : \langle u, v\rangle \in E\} \qquad \text{(outgoing arcs)}, \end{aligned} \tag{1}$$

$|N_G^-(u)|$ is the *in-degree* of u and $|N_G^+(u)|$ is the *out-degree* of u. A sequence $P = \langle u_1, u_2, \ldots, u_n\rangle$ of vertices of G is called a *path* if $\langle u_i, u_{i+1}\rangle \in E$ for $i = 1, 2, \ldots, n-1$, the length of a path P is the number of arcs in the path. We denote by $\mathrm{TS}(G)$ the set of all topological sortings of G. For a topological sorting $\tau \in \mathrm{TS}(G)$ and a subset $E_1 \subseteq E$ we use the following notation:

$$\begin{aligned} \mathcal{T}^{E_1}(\tau) &:= \{\langle v, u\rangle \in E_1 : \tau(u) - \tau(v) = 1\},\\ \mathcal{T}_n^{E_1}(\tau) &:= \{\langle v, u\rangle \in E_1 : \tau(u) - \tau(v) \le n\}. \end{aligned} \tag{2}$$

where n is a positive integer. Given an arc $\langle u, v\rangle \in E$. The number $\tau(v) - \tau(u)$ is called the τ*–distance*.

Let $\pi = \{V_1, V_2, \ldots, V_k\}$ be a partition of V. We denote by $\mathcal{G}(G, \pi)$ the following digraph (directed graph):

$$\langle \pi, \{\langle V_i, V_j\rangle : 1 \le i, j \le k \wedge i \ne j \wedge \underset{u,v\in V}{\exists}(u \in V_i \wedge v \in V_j \wedge \langle u, v\rangle \in E)\}\rangle. \tag{3}$$

Given an undirected graph $G = \langle V, E\rangle$, a vertex cover of G is a subset V' of V such that each edge of E is incident to at least one vertex of V'.

In further considerations a simplified model of proofs is used. The general case of such models that describe proofs written in natural deduction was introduced and considered in [22]. The simplified model of proofs is represented by a DAG $\mathfrak{P}$ with a distinguished set of arcs $\mathfrak{R}(\mathfrak{P})$. The vertices of $\mathfrak{P}$ represent steps of reasoning and arcs of $\mathfrak{P}$ represent the flow of information between different steps of reasoning. Additionally, an arc of $\mathfrak{R}(\mathfrak{P})$ describes the dependence between a step s (the head of the arc) and a previously justified step p (the tail of the arc), called *reference arc*, if the statement formulated in step p is used in the justification of s. Other arcs of $\mathfrak{P}$ describe e.g. the dependence between steps which introduce dummy variables, and steps that contain these variables in the statement.

The methods of improving legibility of proofs based on Behaghel's First Law can be formulated as the following two decision problems:

1st Method of Improving Legibility for n (1st MIL$_n$):

INSTANCE: A DAG $G = \langle V, E\rangle$, a subset E_1 of E and $K \le |E_1|$.

QUESTION: Does there exist a topological sorting τ of G for which $|\mathcal{T}_n^{E_1}(\tau)| \ge K$?

2nd Method of Improving Legibility (2nd MIL):

INSTANCE: A DAG $G = \langle V, E\rangle$, a subset E_1 of E, a positive integer $K \le |V|$.

QUESTION: Does there exist a topological sorting τ of G for which $\tau(u) - \tau(v) \le K$ for every $\langle v, u\rangle \in E_1$?

In our setting, the subset E_1 corresponds to the set of reference arcs. The 1st MIL$_n$ corresponds to the case when the number of local references arcs is optimized. The parameter n that occurs in 1st MIL$_n$ corresponds to the cognitive limit such that references of τ–distance

is less than or equal to n are considered to be comprehensive while the arcs with τ–distance greater – obscure. The 2nd MIL corresponds to the case when we want to construct a topological sorting of the proof graph where every reference arc has τ–distance no greater than the "cognitive limit". Observe that 2nd MIL problem is equivalent to a known NP-complete problem called Directed Bandwidth (see GT41 in [5]). Thus, we immediately conclude that 2nd MIL is NP-complete, too. In our consideration we show that 1st MIL$_n$ problem is also NP-complete.

Having analyzed the "cognitive limit" indicated by the Mizar system users, we can distinguish one subcase of 1st MIL$_n$ problem, where $n = 1$. Before we formulate this case, it should be observed that proofs do not contain self-loops, just like in mathematical proofs it is illegal for a statement of step s to be used in the justification of s. Therefore we can formulate a subcase of 1st MIL$_n$ problem, where $n = 1$ in the following form, that is equivalent to this subcase if a digraph occurring in the instance is without self-loops.

3rd Method of Improving Legibility (**3rd MIL**):

INSTANCE: A DAG $G = \langle V, E\rangle$, a subset E_1 of E, a positive integer $K \leq |V|$.

QUESTION: Does there exist a topological sorting τ of G for which $|\mathcal{T}^{E_1}(\tau)| \geq K$?

This subcase of 1st MIL$_n$ problem presents the case where more comprehensive are only references with premises located directly in the preceding step. Since each step of the reasoning can have at most one premise located in the preceding step, intuitions related to the 3rd MIL problem can be expressed as follows: *a step where at least some of the information it requires is available in the directly preceding step is more comprehensive than a step in which all information is far away in the proof*. This interpretation has a lot in common with the construction `then` implemented in Mizar, Isabelle/Isar and other systems where the proof style inspired by Mizar is implemented: *Declare* [25], *Mizar Mode for HOL* [15], *Mizar-light for HOL-light* [29], *MMode for Coq* [6], *declarative proof language (DPL) for Coq* [4]. The construction `then` indicates that a fact derived directly before should be used in the current step as (part of) its justification. Hence this construction augments the proof context. Additionally, appropriate arrangement of proof steps in a reasoning can increase the number of uses of `then`, and the maximization of this number is realized by the decision problem 3rd MIL.

3 The NP-Completeness of the 1st MIL Problem

It is clear that 1st MIL$_n$ is in the NP class. We can guess a topological sorting τ and count the number of arcs with limited τ–distance. To show that 1st MIL is NP-hard, we transform the Vertex Cover problem, which is known to be NP-complete (see GT41 in [5]) to 3rd MIL. This is done in Theorem 1. Subsequently, we transform 3rd MIL to 1st MIL$_n$ in Theorem 2.

For convenience we recall the Vertex Cover problem.

Vertex Cover (VC):

INSTANCE: An undirected graph $G = \langle V, E\rangle$ and a positive integer $K \leq |V|$.

QUESTION: Is there a vertex cover of size at most K?

Theorem 1 *3rd MIL is NP-complete.*

Proof We transform VC to 3rd MIL. Let an undirected graph $G = \langle V, E\rangle$ and a positive integer $K \leq |V|$ be an instance of VC. We construct a directed graph $G' = \langle V', E'\rangle$ and a subset of arcs $E_1 \subset E'$ such that there exists a vertex cover of G with the size at most K if and only if there exists a topological sorting $\tau \in \mathrm{TS}(G')$ for which $|\mathcal{T}^{E_1}(\tau)| \geq |V| - K$. Let G', E_1 be defined by

$$\begin{aligned} V' &= V \times \{0, 1\}, \\ E' &= \{\langle\langle v, 0\rangle, \langle v, 1\rangle\rangle : v \in V\} \cup \{\langle\langle v, 0\rangle, \langle u, 1\rangle\rangle : \{v, u\} \in E\}, \\ E_1 &= \{\langle\langle v, 0\rangle, \langle v, 1\rangle\rangle : v \in V\}. \end{aligned} \tag{4}$$

This translation can clearly be done in LOGSPACE. Notice that G' is acyclic since for each $v \in V$ the in-degree of vertices $\langle v, 0\rangle$ is 0 and out-degree of vertices $\langle v, 1\rangle$ is 0.

The main idea of the proof is based on the fact that the choice of at least one vertex of every arc $\{v, u\} \in E$ to a vertex cover of G can be expressed by a choice of at least one of arcs $\langle\langle v, 0\rangle, \langle v, 1\rangle\rangle$, $\langle\langle u, 0\rangle, \langle u, 1\rangle\rangle$ to $E_1 \setminus \mathcal{T}^{E_1}(\tau)$. Let us consider an $\{v, u\}$ that belongs to E. First note that the vertices $\langle v, 0\rangle$ and $\langle u, 0\rangle$ are connected with exactly one outgoing arc that belongs to E_1, either $\langle\langle v, 0\rangle, \langle v, 1\rangle\rangle$ or $\langle\langle u, 0\rangle, \langle u, 1\rangle\rangle$. To justify the formulated above fact we show that these arcs must not belong to $\mathcal{T}^{E_1}(\tau)$, for every $\tau \in \mathrm{TS}(G')$ (see Fig. 1). This is a consequence of a simple observation that at most one of two equalities $\tau(\langle v, 0\rangle) + 1 = \tau(\langle v, 1\rangle)$, $\tau(\langle u, 0\rangle) + 1 = \tau(\langle u, 1\rangle)$ can hold for every $\tau \in \mathrm{TS}(G')$. Suppose, contrary to our claim that both hold. Since τ is an injective function, we have $\tau(\langle v, 0\rangle) < \tau(\langle u, 0\rangle)$ or $\tau(\langle u, 0\rangle) < \tau(\langle v, 0\rangle)$. Let us assume that $\tau(\langle v, 0\rangle) < \tau(\langle u, 0\rangle)$ (the other case is analogous). As $\tau(\langle v, 0\rangle) + 1 = \tau(\langle v, 1\rangle)$ and $\langle v, 1\rangle \neq \langle u, 0\rangle$ we have that $\tau(\langle v, 1\rangle) < \tau(\langle u, 0\rangle)$, but this contradicts the fact that $\tau \in \mathrm{TS}(G')$, since $\langle\langle u, 0\rangle, \langle v, 1\rangle\rangle \in E'$.

Let $\mathcal{V}$ be a vertex cover of G with $|\mathcal{V}| \leq K$. Let us consider a partition $\pi(\mathcal{V})$ of G' defined by

$$\pi(\mathcal{V}) := \{\{\langle v, 0\rangle\} : v \in \mathcal{V}\} \cup \{\{\langle v, 1\rangle\} : v \in \mathcal{V}\} \cup \{\{\langle v, 0\rangle, \langle v, 1\rangle\} : v \in V \setminus \mathcal{V}\}. \tag{5}$$

It follows that vertices in $\mathcal{G}(G', \pi(\mathcal{V}))$ have size at most two. Moreover every vertex that has the form $\{\langle v, 0\rangle\}$ has no incoming arcs since $\langle v, 0\rangle$ also does not have, and analogously every vertex that has the form $\{\langle v, 1\rangle\}$ has no outgoing arcs since $\langle v, 1\rangle$ does not have. Hence, cycles can consist only of vertices of size 2 in $\mathcal{G}(G', \pi(\mathcal{V}))$, but vertices of this kind are not connected since $\mathcal{V}$ is a vertex cover. For this reason it is easy to check that $\mathcal{G}(G', \pi(\mathcal{V}))$ is acyclic and, in consequence, there exists $\tau \in \mathrm{TS}(\mathcal{G}(G', \pi(\mathcal{V})))$. Let $\sigma : V' \to \{1, 2, \ldots, |V'|\}$ be the function defined as follows:

$$\sigma(\langle v, i\rangle) = \begin{cases} 1 + \sum\limits_{R \in \pi(\mathcal{V}) : \tau(R) < \tau(P)} |R| & \text{for} \quad |P| = 1, \\ 1 + i + \sum\limits_{R \in \pi(\mathcal{V}) : \tau(R) < \tau(P)} |R| & \text{for} \quad |P| = 2, \end{cases} \tag{6}$$

where P is the only element of $\pi(\mathcal{V})$ that contains $\langle v, i\rangle$. First we show that $\sigma \in \mathrm{TS}(G')$. Indeed, let us consider an arc $\langle\langle w, i\rangle, \langle t, j\rangle\rangle \in E'$ and we denote by P_w, P_t the only element of $\pi(\mathcal{V})$ that contains $\langle w, i\rangle$, $\langle t, j\rangle$, respectively. We have $i = 0$, $j = 1$, because the out-degree of $\langle w, 1\rangle$ is 0 and the in-degree of $\langle t, 0\rangle$ is 0 in G'. Note that if $P_w = P_t$, then $w = t$, $\sigma(\langle w, 0\rangle) + 1 = \sigma(\langle t, 1\rangle)$, and finally $\sigma(\langle w, 0\rangle) < \sigma(\langle t, 1\rangle)$. Suppose that $P_w \neq P_t$. Then $\langle P_w, P_t\rangle$ is an arc in $\mathcal{G}(G', \pi(\mathcal{V}))$, hence $\tau(P_w) < \tau(P_t)$, $\sigma(\langle t, 1\rangle) - \sigma(\langle w, 0\rangle) > \sum\limits_{R \in \pi(\mathcal{V}) : \tau(P_w) < \tau(R) < \tau(P_t)} |R|$, and finally $\sigma(\langle w, 0\rangle) < \sigma(\langle t, 1\rangle)$.

It is also easily seen that $\{\langle\langle v, 0\rangle, \langle v, 1\rangle\rangle \in E_1 : v \in V \setminus \mathcal{V}\} \subseteq \mathcal{T}^{E_1}(\sigma)$, hence finally $\mathcal{T}^{E_1}(\sigma) \geq |V| - K$.

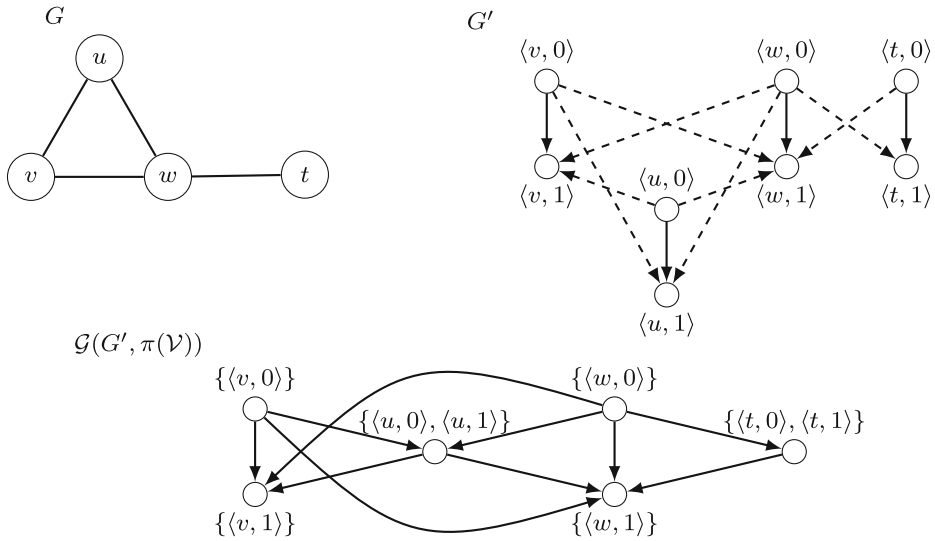

Fig. 1 An example that illustrates the construction from the proof of Theorem 1, where $\mathcal{V} = \{v, w\}$

Let $\sigma \in \mathrm{TS}(G')$, for which $|\mathcal{T}^{E_1}(\sigma)| \geq |V| - K$. Let $\mathcal{V}_\sigma = \{v \in V : \langle\langle v,0\rangle,\langle v,1\rangle\rangle \in E_1 \setminus \mathcal{T}^{E_1}(\sigma)\}$. As $\mathcal{T}^{E_1}(\sigma) \subseteq E_1$, $|E_1| = |V|$ we infer that $|\mathcal{V}_\sigma| \leq K$, so we only need to show that $\mathcal{V}_\sigma$ is a vertex cover of G. Suppose, contrary to our claim, that there exists an edge $\{v,u\} \in E$ such that $\{v,u\} \cap \mathcal{V}_\sigma = \emptyset$. Note that if $\sigma(\langle v,1\rangle) - \sigma(\langle v,0\rangle) > 1$ then $\langle\langle v,0\rangle,\langle v,1\rangle\rangle \notin \mathcal{T}^{E_1}(\sigma)$, and consequently $v \in \mathcal{V}_\sigma$, but this contradicts our assumption that $\{v,u\} \cap \mathcal{V}_\sigma = \emptyset$. Hence $\sigma(\langle v,1\rangle) - \sigma(\langle v,0\rangle) \leq 1$ and analogously $\sigma(\langle u,1\rangle) - \sigma(\langle u,0\rangle) \leq 1$. Additionally, $\sigma(\langle v,0\rangle) < \sigma(\langle u,1\rangle)$, $\sigma(\langle u,0\rangle) < \sigma(\langle u,1\rangle)$, since $\sigma \in \mathrm{TS}(G')$, but these inequalities are between natural numbers, hence finally $\sigma(\langle v,0\rangle) = \sigma(\langle u,0\rangle)$, $\sigma(\langle v,1\rangle) = \sigma(\langle u,1\rangle)$ and $v = u$, but this contradicts our assumption that G is without self-loops. □

Let us take a DAG $G = \langle V, E\rangle$, a subset E' of E and a positive integer n. We define a digraph $G_n = \langle V_n, E_n\rangle$ and a subset E'_n of E_n as:

$$\begin{aligned} V_n &= V \cup \{v_i : v \in V \wedge 1 \leq i \leq n\} \\ E_n &= E \cup \mathcal{E}_{V,n}, \\ E'_n &= E' \cup \mathcal{E}_{V,n} \end{aligned} \tag{7}$$

where $\mathcal{E}_{V,n} := \{\langle v, v_i\rangle : v \in V \wedge 1 \leq i \leq n\}$. Note that the construction of G_n can clearly be done in LOGSPACE. Obviously, G_n is acyclic since for every new vertex $v \in V_n \setminus V$ the out-degree is 0. We show that there exists a topological sorting $\tau \in \mathrm{TS}(G)$ for which $|\mathcal{T}^{E'}(\tau)| \geq K$ if and only if there exists a topological sorting $\sigma \in \mathrm{TS}(G_n)$ for which $|\mathcal{T}^{E'_n}_{n+1}(\sigma)| \geq n \cdot |V| + K$. For this purpose we show firstly that we can rearrange the vertices on every topological sorting of G_n so that all the vertices in V remain in their relative position in a new topological sort of G_n, and the vertices $v_1, v_2, \ldots, v_n$ are all moved right behind v for all $v \in V$.

Lemma 1 *Let E' be a subset of E and σ be a topological sorting of G_n. Then there exists a topological sorting $\sigma' \in TS(G_n)$ such that:*

(i) *if* $v, u \in V$ *and* $\sigma(v) \leq \sigma(u)$, *then* $\sigma'(v) \leq \sigma'(u)$,
(ii) *if* $v \in V$, *then* $\sigma'(v_i) = \sigma'(v) + i$ *for all* $i = 1, 2, \ldots, n$,
(iii) $|\mathcal{T}_{n+1}^{E'_n}(\sigma)| \leq |\mathcal{T}_{n+1}^{E'_n}(\sigma')|$.

Proof We define a sequence $a^1, a^2, \ldots, a^{|V|}$ of the vertices of V based on σ such that $\sigma(a^i) < \sigma(a^j)$ if and only if $i < j$ for each $1 \leq i, j \leq |V|$. We construct a sequence $\Sigma = \langle \sigma_0, \sigma_1, \sigma_2, \ldots, \sigma_{|V|} \rangle$ of topological sortings of G_n based on σ such that $|\mathcal{T}_{n+1}^{E'_n}(\sigma_i)| \geq |\mathcal{T}_{n+1}^{E'_n}(\sigma)|$ for $i = 1, 2, \ldots, |V|$, and then $\sigma' := \sigma_{|V|}$ fulfils the lemma's conditions. The constructed topological sorting rearranges the vertices as follows: $\sigma_0 = \sigma$ and for every $1 \leq i \leq |V|$:

(i) all the vertices in V retain their relative positions σ_i from σ,
(ii) every vertex $v \in V_n$ that occurs before a^i in σ_{i-1} preserves its position in σ_i,
(iii) the vertices $a_1^i, a_2^i, \ldots, a_n^i$ are all moved right behind a^i,
(iv) $|\mathcal{T}_{n+1}^{E'_n}(\sigma)| \leq |\mathcal{T}_{n+1}^{E'_n}(\sigma_k)|$.

Now we proceed by induction over the sorting index number i. We prove the following conditions:

(i) if $v, u \in V_n \setminus \{a_1^i, a_2^i, \ldots a_n^i\}$, $i > 0$, $\sigma_{i-1}(a^i) \leq \sigma_{i-1}(v)$, and $\sigma_{i-1}(v) \leq \sigma_{i-1}(u)$ then $\sigma_i(v) \leq \sigma_i(u)$,
(ii) if $v \in V_n$, $i > 0$, and $\sigma_{i-1}(v) \leq \sigma_{i-1}(a^i)$ then $\sigma_i(v) = \sigma_{i-1}(v)$,
(iii) if $1 \leq j \leq n$, $i > 0$ then $\sigma_i(a_j^i) = \sigma_i(a^i) + j$,
(iv) if $i > 0$ then $|\mathcal{T}_{n+1}^{E'_n}(\sigma_{i-1})| \leq |\mathcal{T}_{n+1}^{E'_n}(\sigma_i)|$.

The case $i = 0$ is obvious. To prove the induction step, assume that we have already constructed σ_i for some i, where $0 \leq i < |V|$. It is clear that there exists $\sigma_{i+1} \in \mathrm{TS}(G')$ that satisfies the following constraints:

(i) if $v, u \in V \setminus \{a_1^{i+1}, a_2^{i+1}, \ldots a_n^{i+1}\}$, $\sigma_i(a^{i+1}) \leq \sigma_i(v)$ and $\sigma_i(v) \leq \sigma_i(u)$ then $\sigma_{i+1}(v) \leq \sigma_{i+1}(u)$,
(ii) if $v \in V_n$ and $\sigma_i(v) \leq \sigma_i(a^{i+1})$ then $\sigma_{i+1}(v) = \sigma_i(v)$.
(iii) if $1 \leq j \leq n$ then $\sigma_{i+1}(a_j^{i+1}) = \sigma_i(a^{i+1}) + j$.

To finish the proof we show that $|\mathcal{T}_{n+1}^{E'_n}(\sigma_i)| \leq |\mathcal{T}_{n+1}^{E'_n}(\sigma_{i+1})|$. It is easy to check that to compare the size of the sets it is enough to check the number of outgoing arcs from a^{i+1} in both these sets. Suppose contrary to our claim, that $\mathcal{T}_{n+1}^{E'_n}(\sigma_i)$ has more than $\mathcal{T}_{n+1}^{E'_n}(\sigma_{i+1})$ such kind of arcs. From (2) we infer that $\mathcal{T}_{n+1}^{E'_n}(\sigma_i)$ cannot have more than $n + 1$ such arcs. But from (iii) we conclude that $\langle a^{i+1}, a_j^{i+1} \rangle \in \mathcal{T}_{n+1}^{E'_n}(\sigma_{i+1})$ for $j = 1, 2, \ldots, n$, hence $\mathcal{T}_{n+1}^{E'_n}(\sigma_i)$ has to have exactly $n + 1$ arcs outgoing from a^{i+1}. Since not all these arcs can be in the form $\langle a^{i+1}, a_j^{i+1} \rangle$ for $j = 1, 2, \ldots, n$, there exists $v \in V$ such that $\langle a^{i+1}, v \rangle \in \mathcal{T}_{n+1}^{E'_n}(\sigma_i)$. Additionally, we can assume that $\sigma_i(v)$ has the smallest value among all $\sigma_i(u)$ such that $u \in V$ and $\langle a^{i+1}, u \rangle \in \mathcal{T}_{n+1}^{E'_n}(\sigma_i)$. Clearly, $\sigma_{i+1}(v) = \sigma_{i+1}(a^{i+1}) + n + 1$, hence $\langle a^{i+1}, v \rangle \in \mathcal{T}_{n+1}^{E'_n}(\sigma_{i+1})$ and consequently, $\mathcal{T}_{n+1}^{E'_n}(\sigma_{i+1})$ has the same outgoing arcs from a^{i+1} as $\mathcal{T}_{n+1}^{E'_n}(\sigma_i)$, a contradiction. □

Lemma 2 *Let E' be a subset of E and σ be a topological sorting of G_n. Then there exists a topological sorting $\tau \in TS(G)$ such that*

$$|\mathcal{T}^{E'}(\tau)| \geq |\mathcal{T}_{n+1}^{E'_n}(\sigma)| - n \cdot |V|.$$

Proof Let $\sigma \in \mathrm{TS}(G_n)$. From Lemma 1 we conclude that there exists $\sigma' \in \mathrm{TS}(G_n)$ such that

$$\sigma'(v_i) = \sigma'(v) + i \tag{8}$$

for every $v \in V$, $i = 1, 2, \ldots, n$, and $|\mathcal{T}_{n+1}^{E'_n}(\sigma)| \leq |\mathcal{T}_{n+1}^{E'_n}(\sigma')|$. Define a sequence $a^1, a^2, \ldots, a^{|V|}$ of the vertices of V as follows: $\sigma'(a^i) < \sigma'(a^j)$ if and only if $i < j$ for all $1 \leq i, j \leq |V|$, and let $\tau : V \to \{1, 2, \ldots, |V|\}$ be the function given by the formula $\tau(a^i) = i$ for each $i = 1, 2, \ldots, |V|$. Note that τ is a topological sorting of G since $\sigma' \in \mathrm{TS}(G_n)$. Hence to complete the proof it is enough to show that $|\mathcal{T}^{E'}(\tau)| = |\mathcal{T}_{n+1}^{E'_n}(\sigma')| - n \cdot |V|$, since $|\mathcal{T}_{n+1}^{E'_n}(\sigma')| - n \cdot |V| \geq |\mathcal{T}_{n+1}^{E'_n}(\sigma)| - n \cdot |V|$. For this purpose we show only that $\mathcal{T}_{n+1}^{E'_n}(\sigma') = \mathcal{T}^{E'}(\tau) \cup \mathcal{E}_{V,n}$, since it is evident that $\mathcal{T}^{E'}(\tau) \cap \mathcal{E}_{V,n} = \emptyset$ and $|\mathcal{E}_{V,n}| = n \cdot |V|$.

Observe that $\mathcal{E}_{V,n} \subseteq \mathcal{T}_{n+1}^{E'_n}(\sigma')$, which follows from (8). To show that $\mathcal{T}^{E'}(\tau) \subseteq \mathcal{T}_{n+1}^{E'_n}(\sigma')$ let us consider $\langle v, u \rangle \in \mathcal{T}^{E'}(\tau)$. Then there exists $1 \leq i \leq |V| - 1$ such that $v = a^i$, $u = a^{i+1}$, since $\tau(u) - \tau(v) = 1$. Additionally, all vertices that have the form a^i_j for $j = 1, 2, \ldots, n$ and only these vertices are located between a^i and a^{i+1} in the linear arrangement σ'. Hence $\sigma'(u) - \sigma'(v) = n + 1$ and finally $\langle v, u \rangle \in \mathcal{T}_{n+1}^{E'_n}(\sigma')$.

To show the last inclusion $\mathcal{T}_{n+1}^{E'_n}(\sigma') \subseteq \mathcal{T}^{E'}(\tau) \cup \mathcal{E}_{V,n}$, let us consider an arc $\langle w, t \rangle \in \mathcal{T}_{n+1}^{E'_n}(\sigma')$ and assume that $\langle w, t \rangle \notin \mathcal{E}_{V,n}$. By assumption, $\langle w, t \rangle \in E'_n$, hence $\langle w, t \rangle \in E'$ since $E'_n = E' \cup \mathcal{E}_{V,n}$ and $\langle w, t \rangle \notin \mathcal{E}_{V,n}$. Consequently, $w, t \in V$, but between every two different vertices of V in the linear arrangement σ' at least n vertices belong to $V_n \setminus V$, hence $\sigma'(t) - \sigma'(w) \geq n + 1$. Additionally $\sigma'(t) - \sigma'(w) \leq n + 1$, since $\langle w, t \rangle \in \mathcal{T}_{n+1}^{E'_n}(\sigma')$. It follows that between w, t in σ' there are no vertices that belong to V, hence there exists j such that $w = a^j$, $t = a^{j+1}$, $1 \leq j \leq |V| - 1$, and finally $\langle w, t \rangle \in \mathcal{T}^{E'}(\tau)$, since $\langle w, t \rangle \in E'$. □

This result can be strengthened to evidence that giving more freedom to put premises farther from their place of use cannot give us any polynomial algorithm.

Theorem 2 *The problem 1st MIL$_n$ is NP-complete for each n.*

Proof The case $n = 1$ is obvious since we proved in Theorem 1 that 3rd MIL is NP-complete problem. Assume that $n > 1$. For this case we transform 3rd MIL to 1st MIL$_n$. Let a DAG $G = \langle V, E \rangle$, E' be a subset of E and a positive integer $K \leq |V|$ be an instance of 3rd MIL. Recall the definition (7) of graph G_n in the proof of Theorem 1. We prove that there exists a topological sorting $\tau \in \mathrm{TS}(G)$ for which $|\mathcal{T}^{E'}(\tau)| \geq K$ if and only if there exists a topological sorting $\sigma \in \mathrm{TS}(G_{n-1})$ for which $|\mathcal{T}_n^{E'_{n-1}}(\sigma)| \geq (n-1) \cdot |V| + K$.

Let τ be a topological sorting of G for which $|\mathcal{T}^{E'}(\tau)| \geq K$, and let us consider the function $\sigma : V_{n-1} \to \{1, 2, \ldots, |V_{n-1}|\}$ defined as follows:

$$\sigma(v) = 1 + n \cdot (\tau(v) - 1), \qquad \sigma(v_i) = 1 + i + n \cdot (\tau(v) - 1), \tag{9}$$

for every $v \in V$ and $i = 1, 2, \ldots, n-1$. Observe that $\sigma \in \mathrm{TS}(G_{n-1})$, since $\sigma(v) < \sigma(u)$ for each $\langle v, u\rangle \in E$, and $\sigma(v) < \sigma(v_i)$ for each $v \in V$, $1 \le i \le n-1$. By the definition of σ we obtain that $\mathcal{T}^{E_1}(\tau) \cup \mathcal{E}_{V,n-1} \subseteq \mathcal{T}_n^{E'_{n-1}}(\sigma)$, hence finally $|\mathcal{T}_n^{E'_{n-1}}(\sigma)| \ge (n-1)\cdot|V| + K$.

Let σ be a topological sorting of G_{n-1} for which $|\mathcal{T}_n^{E'_{n-1}}(\sigma)| \ge (n-1)\cdot|V| + K$. By Lemma 2 we infer that there exists $\tau \in \mathrm{TS}(G)$ such that $|\mathcal{T}^{E'}(\tau)| \ge |\mathcal{T}_n^{E'_{n-1}}(\sigma)| - (n-1)\cdot|V|$ since $n > 1$. Hence finally $|\mathcal{T}^{E'}(\tau)| \ge K$. □

Analyzing the family of graphs constructed in Theorem 1, 2 we obtain that 1st MIL_n is NP-complete only for instances, where subsets E, E_1 are not equal. The case $E = E_1$ is also NP-complete, and it was proved in [24], but the transformation there is not in LOGSPACE, and we do not know how to improve it to be in LOGSPACE.

4 Problem 3rd MIL as a Finding of a Hamiltonian Path

In this section we state and classify an unexplored problem equivalent to 3rd MIL that will be determined in the terms of Hamiltonian paths for acyclic digraph. Such a statement enables interpreting the 3rd MIL problem among the problems known as Hamiltonian Completion problems that correspond to the existence of a Hamiltonian path.

Directed Hamiltonian Path Completion (**Directed HPC**):

INSTANCE: A DAG $G = \langle V, E\rangle$, a subset E_1 of E, a positive integer K.

QUESTION: Does there exist E'_1 being a subset of V^2 containing E_1, such that $|E'_1 \setminus E_1| \le K$ and the digraph $\langle V, E \cup E'_1\rangle$ is acyclic and the digraph $\langle V, E'_1\rangle$ has a Hamiltonian path?

Note that in the reality of graphs undirected completion to a Hamiltonian circuit (see GT34 in [5]) or path (see GT39 in [5]) are known NP-complete problems. Additionally, Directed Hamiltonian Circuit (see GT38 [5]), Directed Hamiltonian Path [5] are also NP-complete for digraphs, but the second problem can be solved in polynomial time for acyclic digraphs [18].

In this section we show in Theorem 4 that Directed HCP is also NP-complete, since 3rd MIL is NP-complete too. Additionally, as 3rd MIL is NP-complete even for $E = E_1$ (see [24]), we obtain by Theorem 3 that E_1 can be replaced by E in instances of Directed HCP.

Observe first that K occurring in the instance of Directed HCP can be restricted by $|V| - 1$, since for $K \ge |V| - 1$ a solution always exists. Indeed, let τ be a topological sorting of G, then it is evident that $E'_1 = E_1 \cup \mathcal{E}(\tau)$ is a solution of Directed HCP, where $\mathcal{E}(\tau) := \{\langle \tau^{-1}(i), \tau^{-1}(i+1)\rangle : 1 \le i \le |V| - 1\}$.

Theorem 3 *Let $G = \langle V, E\rangle$ be a DAG, E_1 be a subset of E and K be a positive integer not greater than $|V| - 1$. The following conditions are equivalent:*

(i) there exists a topological sorting τ of G for which $|\mathcal{T}^{E_1}(\tau)| \ge K$,
(ii) there exists E'_1 be a subset of V^2 containing E_1, such that $|E'_1 \setminus E_1| \le |V| - K - 1$, the digraph $\langle V, E \cup E'_1\rangle$ is acyclic and the digraph $\langle V, E'_1\rangle$ has a Hamiltonian path.

Proof $(i) \Longrightarrow (ii)$ Let $\tau \in \mathrm{TS}(G)$ for which $|\mathcal{T}^{E_1}(\tau)| \ge K$. Let us consider $E'_1 := E_1 \cup \mathcal{E}(\tau)$. Observe that $\mathcal{T}^{E_1}(\tau) \subseteq \mathcal{E}(\tau)$, hence $E'_1 \setminus E_1 \subseteq \mathcal{E}(\tau) \setminus \mathcal{T}^{E_1}(\tau)$, and finally $|E'_1 \setminus E_1| \le$

$|V| - 1 - K$. Additionally $\tau \in \mathrm{TS}(\langle V, E \cup E_1' \rangle)$ and $\langle \tau^{-1}(1), \tau^{-1}(2), \ldots, \tau^{-1}(|V|) \rangle$ is a Hamiltonian path of $\langle V, E_1' \rangle$, hence finally the proof of the first implication is complete.

$(ii) \implies (i)$ Let us consider E_1' being a subset of V^2 containing E_1, such that $|E_1' \setminus E_1| \leq |V| - K - 1$, the digraph $\langle V, E \cup E_1' \rangle$ is acyclic, and $\langle V, E_1' \rangle$ has a Hamiltonian path $\mathfrak{h} = \langle v_1, v_2, \ldots, v_{|V|} \rangle$. Let $\tau : V \to \{1, 2, \ldots, |V|\}$ be a function given by the formula $\tau(v_i) = i$ for each $i = 1, 2, \ldots, |V|$. We claim that $\tau \in \mathrm{TS}(G)$. Indeed, suppose contrary to our claim, that there exists an arc $\langle u, w \rangle \in E$ such that $\tau(u) > \tau(w)$. Then $\langle w, v_{\tau(w)+1}, v_{\tau(w)+2}, \ldots, v_{\tau(u)-1}, u, w \rangle$ is a cycle, but this contradicts our assumption that $\langle V, E \cup E_1' \rangle$ is acyclic. Since $\tau \in \mathrm{TS}(G)$ we need to show only that $|\{\langle v, u \rangle \in E_1 : \tau(u) - \tau(v) = 1\}| \leq K$. Note that $\mathcal{E}(\tau) \subseteq E_1'$ since $\mathfrak{h}$ is a path of $\langle V, E_1' \rangle$. Additionally $\mathcal{E}(\tau) \setminus \mathcal{T}^{E_1}(\tau) = \mathcal{E}(\tau) \setminus E_1$ and $\mathcal{E}(\tau) \setminus E_1 \subseteq E_1' \setminus E_1$, hence $|\mathcal{E}(\tau) \setminus \mathcal{T}^{E_1}(\tau)| \leq |V| - K - 1$. As $|\mathcal{E}(\tau)| = |V| - 1$, $\mathcal{T}^{E_1}(\tau) \subseteq \mathcal{E}(\tau)$ we have that $|\mathcal{T}^{E_1}(\tau)| \geq K$. □

An easy computation shows that from Theorem 1 and 3 we can infer the following theorem.

Theorem 4 *Directed HPC is NP-complete.*

Note that searching of the set $E_1' \setminus E_1$ can be limited to a subset of $V^2 \setminus E_1$. As $\langle V, E \cup E_1' \rangle$ is acyclic, we obtain that $E_1' \setminus E$ cannot contain any arc $\langle u, v \rangle$ that generates a circle in the digraph $\langle V, E \cup \{\langle u, v \rangle\} \rangle$. Additionally, if there exists a directed path of $\langle V, E \rangle$ that leads from u to v and has length at last 2 then $\langle V, E_1' \rangle$ has a Hamiltonian path if and only if $\langle V, E_1' \setminus \{\langle u, v \rangle\} \rangle$ also does, hence such kind of arcs may also be ignored in the search of $E_1' \setminus E_1$.

5 Conclusions

We concentrated on two methods of improving proof readability based on Behaghel's First Law. We proved that the most comprehensive interpretation of this law lead to optimization of the problems corresponding with NP-complete decision problems. Additionally, for one of these problems we have found an equivalent formulation as unexplored problem Directed-HCP concerning the existence of Hamiltonian paths in digraphs. This problem enables interpreting considered proof readability improvement problems among NP-complete problems of finding a Hamiltonian path.

From the point of view of future authors of formal proofs it is also important to find heuristics that quickly give satisfactory approximate solutions to both problems: 1st and 2nd MIL in more appropriate hierarchy of these problems. The next step in the process of improving proof readability should be finding algorithms that approximate the problem 1st MIL and its subcase 3rd MIL. The successful application of SMT technology to solving computationally difficult problems suggests that application of SMT solvers can also be an effective way of finding solutions to MIL problems.

References

1. Alama, J., Kohlhase, M., Mamane, L., Naumowicz, A., Rudnicki, P., Urban, J.: Licensing the Mizar Mathematical Library. In: Davenport, J.H., Farmer, W.M., Urban, J., Rabe, F. (eds.) Proceedings of the 18th Calculemus and 10th International Conference on Intelligent Computer Mathematics Lecture Notes in Computer Science, vol. 6824, pp. 149–163. Springer-Verlag, Berlin, Heidelberg (2011)
2. Bancerek, G.: The fundamental properties of natural numbers. Formalized Math. **1**(1), 41–46 (1990)
3. Behaghel, O.: Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. Indogermanische Forschungen **25**, 110–142 (1909)
4. Corbineau, P.: A Declarative Language for the Coq Proof Assistant. In: Proceedings of the 2007 International Conference on Types for Proofs and Programs, pp. 69–84 (2007)
5. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. A Series of Books in the Mathematical Science. W. H. Freeman and Company, New York (1979)
6. Giero, M., Wiedijk, F.: MMode, A Mizar Mode for the proof assistant Coq. Technical report, ICIS. Radboud Universiteit Nijmegen (2004)
7. Gonthier, G.: Formal Proof—The Four-Color Theorem. Notices of the AMS **55**(11), 1382–1393 (2008)
8. Grabowski, A., Korniłowicz, A., Naumowicz, A.: Mizar in a Nutshell. J. Formalized Reason. **3**(2), 153–245 (2010)
9. Grabowski, A., Schwarzweller, C.: Translating mathematical vernacular into knowledge repositories. In: Proceedings of the 4th international conference on Mathematical Knowledge Management MKM'05, pp. 49–64. Springer-Verlag, Berlin, Heidelberg (2006)
10. Grabowski, A., Schwarzweller, C.: Revisions as an Essential Tool to Maintain Mathematical Repositories. In: Proceedings of the 14th symposium on Towards Mechanized Mathematical Assistants: 6th International Conference, Lecture Notes in Computer Science, vol. 4573, pp. 235–249. Springer-Verlag (2007)
11. Grabowski, A.: Automated discovery of properties of rough sets. Fundamenta Informaticae **128**(1-2), 65–79 (2013)
12. Grabowski, A.: Efficient rough set theory merging. Fundamenta Informaticae **135**(4), 371–385 (2014)
13. Grabowski, A., Jastrzębska, M.: A note on a formal approach to rough operators. In: Szczuka, M.S., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) Rough Sets and Current Trends in Computing - 7th International Conference, RSCTC 2010, Warsaw, Poland, June 28-30,2010. Proceedings, volume 6086 of Lecture Notes in Computer Science, pp. 307–316. Springer (2010)
14. Grabowski, A., Schwarzweller, C.: Towards automatically categorizing mathematical knowledge. In: Ganzha, M., Maciaszek, L.A., Paprzycki, M. (eds.) Federated Conference on Computer Science and Information Systems – FedCSIS 2012, Wroclaw, Poland, 9–12 September 2012, Proceedings, pp. 63–68 (2012)
15. Harrison, J.: A Mizar Mode for HOL. In: Proc. of the 9th International Conference on Theorem Proving in Higher Order Logics, pp. 203–220. Springer (1996)
16. Korniłowicz, A.: Tentative Experiments with Ellipsis in Mizar. In: Intelligent Computer Mathematics Lecture Notes in Computer Science, vol. 7362, pp. 453–457 (2012)
17. Korniłowicz, A.: On Rewriting Rules in Mizar. J. Autom. Reason. **50**(2), 203–210 (2013)
18. Lawler, E.L.: Combinatorial Optymization: Networks and Matroids. Holt, Rinehart and Winston (1967)
19. Naumowicz, A., Byliński, C.: Improving MIZAR texts with properties and requirements. In: Asperti, A., Bancerek, G., Trybulec, A. (eds.) Third International Conference Mathematical Knowledge Management 2004, MKM'04 Lecture Notes in Computer Science, vol. 3119, pp. 290–301. Springer-Verlag (2004)
20. Naumowicz, A., Korniłowicz, A.: A Brief Overview of Mizar. In: TPHOLs'09, Lecture Notes in Computer Science, vol. 5674, pp. 67–72. Springer-Verlag (2009)
21. Naumowicz, A.: Interfacing external CA systems for Grobner bases computation in Mizar proof checking. Int. J. Comput. Math. **87**(1), 1–11 (2010)
22. Pąk, K.: The algorithms for improving and reorganizing natural deduction proofs. Stud. Logic Grammar Rhetoric **22**(35), 95–112 (2010)
23. Pąk, K.: Methods of lemma extraction in natural deduction proofs. J. Autom. Reason. **50**(2), 217–228 (2013)
24. Pąk, K.: Improving legibility of natural deduction proofs is not trivial. Logical Methods in Computer Science **10**(3:23), 1–30 (2014)
25. Syme, D.: Three Tactic Theorem Proving. In: Theorem Proving in Higher Order Logics, Lecture Notes in Computer Science, vol. 1690, pp. 203–220. Springer-Verlag (1999)
26. Trybulec, A., Korniłowicz, A., Naumowicz, A., Kuperberg, K.: Formal mathematics for mathematicians. J. Autom. Reason. **50**(2), 119–121 (2013)

27. Urban, J.: XML-izing Mizar: Making Semantic Processing and Presentation of MML Easy. In: Bonacina, M.P. (ed.) 4th International Conference Mathematical Knowledge Management 2005, MKM'05 volume 3863 of Lecture Notes in Computer Science, pp. 346–360. Springer-Verlag (2005)
28. Wenzel, M.: The Isabelle/Isar Reference Manual. University of Cambridge (2011)
29. Wiedijk, F.: Mizar Light for HOL Light. In: Proceedings of the 14th International Conference on Theorem Proving in Higher Order Logics, pp. 378–394 (2001)