Utilizing compression and refinement to handle large cases in crime analysis

Susan W. van den Braak[1] and **Herre van Oostendorp**[1] and **Gerard A.W. Vreeswijk**[1] and **Henry Prakken**[2]

Abstract. When large arguments are produced graph visualizations are often hard to read. Argument visualization software should therefore offer features that allow users to display their graphs in a readable way. More specialized software for crime analysts should also offer the ability to elaborate on graphs and to hide redundant nodes. While doing so, it should be easy to unfold all hidden information about a certain node if desired. Therefore, refinement and compression methods are implemented in the *AVERs* software for crime analysts. This paper presents the results of a study that tested the effect of compression and refinement on the quality of the users' analysis of a simple crime case and their understanding of this case. In this study professional crime analysts and students who used these methods outperformed users that were only allowed to use conventional methods to handle large graphs.

1 Introduction

Graph visualization tools have been shown promising for learning and for collaborating on the construction and evaluation of arguments [6, 15]. Increasingly, such tools are applied to the domain of reasoning about evidence in legal settings [19, 8, 18]. They are claimed to support crime analysts who are faced with large quantities of data in structuring and keeping track of the data. Such tools allow them to visualize their reasoning in a way that is meaningful to them and explore its consequences. As a result it should become easier to pinpoint possible gaps and inconsistencies, and strong and weak points in their arguments. In this spirit, a tool for the visualization of stories and evidence named *AVERs* (Argument Visualization for Evidential Reasoning based on stories) has been developed [17, 3]. This tool combines such sense-making approaches, which focus on argumentation, with a story-based approach inspired by legal psychology [10, 20].

In the current practice of crime analysts, software for managing and visualizing evidence is already being used that allows them to formulate stories as simple timelines. A well-known example is Analyst's Notebook [1]. However, a limitation of such software is that it does not allow for expressing the reasons why certain pieces of evidence support or attack a certain hypothesis. Using *AVERs* analysts can do both: they can construct possible stories about what happened by linking events through causal connections and they can connect the available evidence with these stories through arguments (i.e. evidential links). In this way an important aspect of the current practice in Dutch police regions is covered, namely that crime investigators and analysts turn to the reconstruction of what might have happened into stories [4]. In addition, this tool allows analysts to connect their story to the available evidence and thus to represent how this evidence supports or attacks their hypotheses. In this way, the reasons why certain pieces of evidence support or attack a story are made explicit, and hopefully biases are reduced.

The *AVERs* system is based on a theoretical model of reasoning about evidence [3]. In this reasoning model the construction of stories to explain the available evidence is modeled as abductive reasoning with networks of causal generalizations, while source-based reasoning about evidence is modeled as default reasoning with evidential generalizations. The system thus supports both the construction of stories and the construction of arguments and counterarguments. This formal model concerns more than just the construction of visualizations, it also accounts for the comparison of stories. In this way *AVERs* allows its users to compare stories and to critically examine them, for example, by allowing users to reason about the links in a story. This is important since typically an investigation will result in more than one possible hypothesis of what might have happened and an analyst has to choose the best explanation.

Although visualization approaches to reasoning about legal evidence seem promising, and are often claimed to be beneficial, experiments that investigate the effects of visualization tools on the users' reasoning skills are relatively sparse. The little research that has been conducted did not produce clear empirical results, as most of the conducted studies were not valid or failed to show significant effects [15]. Moreover, these experiments concentrated on more general reasoning skills and conflict resolution skills, and as far as we know such studies were not conducted to measure the effect of argument visualization software on the task of crime analysts. This paper will present the results of a study that investigates the effectiveness of the *AVERs* system, which is aimed at crime analysts.

Furthermore, an important problem of structuring data into graphs is that as soon as the size of the graph or the link density increases, such graphs become increasingly more complex and harder to understand. Large graphs will ask a lot of the users' cognitive abilities, as it will become more difficult for them to read and interact with these graphs. For instance, they have to keep overview of the complete graph, keep track of changes, and quickly find the information they are looking for without making mistakes. Readability of large graphs may be enhanced in a few simple ways, for example by scaling (fitting to screen) or flipping graphs (i.e. arranging graphs from top to bottom instead of from left to right). Such simple features allow for graphs to be displayed in such a way that they fit a normal computer screen. Allowing users to zoom in on specific parts of the graph results in temporarily smaller (i.e. a smaller number of nodes is displayed) and better readable graphs (i.e. the nodes that are dis-

[1] Department of Information and Computing Sciences, Utrecht University, email: susanb,herre,gv@cs.uu.nl

[2] Department of Information and Computing Sciences, Utrecht University & Faculty of Law, University of Groningen, email: henry@cs.uu.nl

played are bigger). In this way, the amount of information that is presented to the user at the same time is smaller compared to the situation in which he is confronted with the complete graph, so that readability is improved.

However, there are reasons to believe that such simple features are insufficient for the targeted users of *AVERs*, being crime analysts. In meetings with analysts who are working on large cases on a regular basis, the wish for methods which allow them to maintain overview of the complete story, while they are also able to focus on smaller details has been expressed. This desire has several reasons. Firstly, due to the interactive nature of crime investigations, stories are continuously being refined. More specifically, at first stories are general, but as the investigation unfolds they will become more and more detailed, as more and more information becomes available. Useful software should therefore allow analysts to elaborate on their stories. It is especially important to refine a certain part of a story when questions about the truth of it arise. Secondly, reasons why certain links were established are often left implicit, but sometimes it is necessary to make these reasons explicit when questions about their validity arise. Therefore, features are necessary that allow analysts to zoom in on an element of a story, explore its status, and explain it in more detail or question its truth if necessary, but also to keep overview of the larger picture. This means that graphs should be scaled to make them readable by hiding redundant nodes (those that are detailed and not necessary to understand the main story or argument), but that it should be easy to unfold all hidden information about a certain node if desired. For systems to be usable for crime analysts they should provide different abstraction layers; an overview (summary) level and a node level.

We suggest that a combination of compression and refinement is a viable way to provide these layers, while it also improves the readability of graphs. These methods are inspired by the compression rationale [7], which is based on the idea that sometimes lines of reasoning are compressed into a rule. Take for example, the two-step argument for believing witnesses: as a rule "witnesses that speak the truth should be believed" and "witnesses normally speak the truth". This may then be compressed into the rule "witnesses should be believed". If a rule is to be attacked, it has to be restated in an uncompressed form. It is then easy to see that this rule can be attacked by arguing that witnesses who have a reason to lie do not speak the truth and therefore should not be believed. Compression of links is proposed for our system as it keeps the graph manageable. We distinguish between two actions that may be performed on compressed links. Firstly, the decompression of these compressed links if necessary. This is used in order to be able to add reasons for questionable links, to add counterarguments, and to make the underlying reasoning explicit. Consider for example the rule "if a witness testifies that P is the case then usually P is the case" which may be attacked by saying that the witness is not truthful. Secondly, the refinement of links, that is, the replacement of a link by a chain of links with the same start and end point as the original one. This allows for the addition of more detail to earlier established links. For example, the rule "if a witness testifies that he saw P then usually P is the case" may be refined into "if a witness testifies that he saw P then usually he saw P" and "if a witness saw P then usually P is the case". Useful software tools for crime analysts should thus provide a combination of these two methods, that is, (de)compression and refinement, since these correspond to tasks that are necessary and important during the analysis of a case. In the remainder of this paper, with compression we mean both the action of compressing and decompressing links.

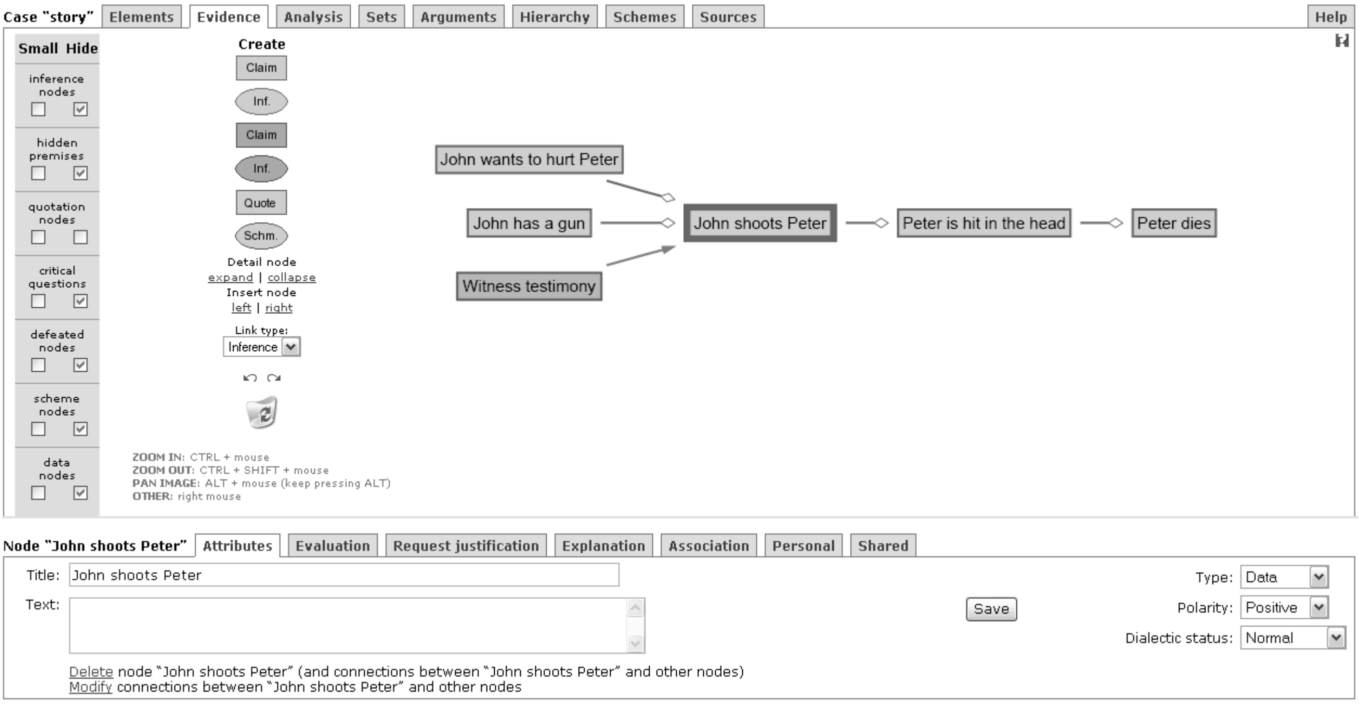
Although compression and refinement seem promising and useful, some state-of-the-art sense-making tools, such as Araucaria [11] and ATHENA [12], only allow for zooming in and out on large graphs and do not provide more advanced methods to make data more manageable. Besides, to our knowledge no research is done on the effect of compression and refinement on the users' performance. This paper reports on a study that was conducted to determine whether tools that contain such methods support their users better than tools that only provide standard methods to handle large graphs. We have done this by measuring the effect of compression and refinement on the quality of crime analyses. The results obtained through this study are also relevant for similar argument visualization tools that display arguments as graphs.

1.1 The *AVERs* system

In the formal model underlying our system we distinguish between causal and evidential generalizations. Causal generalizations represent causal knowledge from cause to effect as in "fire causes smoke" or "if somebody shoots a person, this can cause that person to die". Evidential generalizations are effect-to-cause statements, for instance, "smoke means fire" or "if a witness testifies that an event happened, this is evidence for the occurrence of that event". In our framework, stories are modeled as networks of causal links, while the links between stories and arguments are modeled as evidential links. The *AVERs* software can thus be used to draw such evidential arguments and causal networks. It consists of a split screen where the upper half displays a global overview of the case (the argument graph containing nodes and links) and the lower half displays the attributes of a selected node (see Figure 1). New nodes can be added to the screen by clicking the desired node type and two nodes can be connected by drawing lines from node to node. Thus, a case is built. The graph visualizations in *AVERs* make extensive use of colors, which cannot be shown here. Therefore, in the figures presented in this paper color indications are provided between square brackets. A version with color pictures has been put online at `http://people.cs.uu.nl/susanb/publications/cmna08.pdf`.

In Figure 2 a simple story is displayed. Such a story can be constructed by linking claims about a case (events), represented as green boxes, through causal links which are yellow with diamond-shaped arrowheads. Let us suppose that in this example case the investigation started with the observation that Peter was hit in the head by a bullet and died. The analyst will thus start his analysis by adding two new nodes to the case, `Peter is hit in the head` and `Peter dies`, and connecting them through a causal link. As an explanation for the fact that Peter was hit in the head the police assumed that John wanted to hurt Peter and shot him. More specifically, the observation that John wanted to hurt Peter together with the fact that he had a gun caused him to shoot Peter. The investigators may now add these events that are not yet supported by evidence as nodes to the graph and link them through causal links. This results in a causal chain from `John wants to hurt Peter` and `John has a gun` to `Peter dies`.

Evidence may be added to the graph by selecting text from source documents; such quotes are represented as blue boxes. Stories may be connected with the available evidence through evidential links that are represented as blue arrows. These evidential links are instantiations of argumentation schemes that represent predefined patterns of reasoning that often occur in evidential reasoning [21, 2]. Such schemes contain a conclusion, one or more premises and several critical questions. In the sample story, a witness Jane declared that she saw John shooting at Peter. The analyst may now use the argumenta-



ZOOM IN: CTRL + mouse
ZOOM OUT: CTRL + SHIFT + mouse
PAN IMAGE: ALT + mouse (keep pressing ALT)
OTHER: right mouse

Title: John shoots Peter
Text:
Type: Data
Polarity: Positive
Dialectic status: Normal

Delete node "John shoots Peter" (and connections between "John shoots Peter" and other nodes)
Modify connections between "John shoots Peter" and other nodes

Figure 1. A screenshot of the *AVERs* software

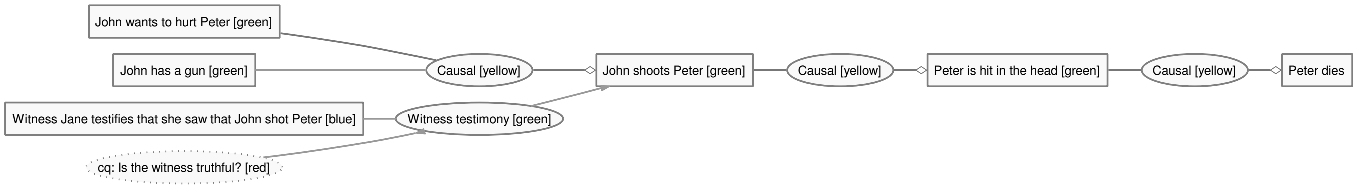
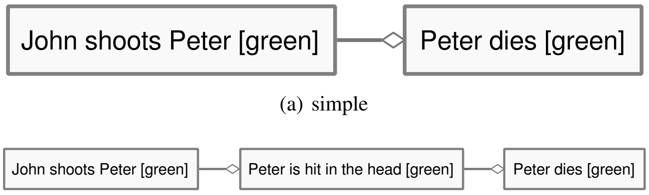

John wants to hurt Peter [green]
John has a gun [green]
Witness Jane testifies that she saw that John shot Peter [blue]
cq: Is the witness truthful? [red]
Causal [yellow]
Witness testimony [green]
John shoots Peter [green]
Causal [yellow]
Peter is hit in the head [green]
Causal [yellow]
Peter dies

Figure 2. Graph visualization of a simple story in *AVERs*

tion scheme for witness testimonies to connect the quote to the corresponding event in his story. This mechanism works as follows, given the rule that witnesses usually tell the truth the event that John actually shot Peter may be inferred and an evidential link between source and event is established. However, while applying such a scheme the user has to take the critical question of whether the witness is truthful into account. A negative answer to this question invalidates the application of the scheme and is therefore added to the scheme as a defeater, which is colored red (in the example a latent defeater is added since the question was answered positively, i.e. witness Jane is believed to be truthful). Scheme instantiations are represented as elipses that display the type of the scheme that was used (also referred to as inference nodes). In this way, such schemes provide justifications for the established links, that is, reasons why such links were created. For a more elaborate explanation of the data model and the system's functionality we refer to [16, 17].

1.2 Compression and refinement

Refinement and compression methods are incorporated into the *AVERs* system to allow analysts to handle large graphs while they are also able to elaborate on their stories and make their underlying reasoning explicit. Firstly, refinement allows for the addition of a new node in between two previously connected nodes to refine stories and add more detail to links. Consider for example Figure 3(a) and suppose that at first the analyst created a link between `John shoots Peter` and `Peter dies`. When information becomes available that the victim died because of a gunshot wound to his head, the analyst may want to specify this link and add the node `Peter is hit in the head` (see Figure 3(b)).

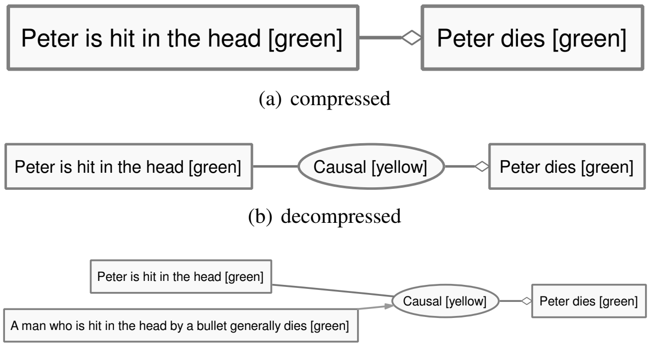

John shoots Peter [green]
Peter dies [green]
John shoots Peter [green]
Peter is hit in the head [green]
Peter dies [green]

(a) simple

(b) refined

Figure 3. Refinement of a causal link in *AVERs*

Secondly, compression involves the folding and unfolding of redundant nodes and links. By default all links are compressed, but users may expand them if they want to add support or add defeaters to attack them. In this way decompression points out possibilities for attack and allows for underlying reasons to be made explicit. For example, in Figure 4(b) the link between `Peter is hit in the head` and `Peter dies` from Figure 4(a) is expanded to reveal the causal nature of the link. The analyst may now decide to add a reason for this link, for example that people who are hit in the head generally die because of this (see Figure 4(c)). Similarly, compression may be used to attack links; this is what is done automatically when critical questions are answered negatively (see Figure 2). For example, in Figure 5(b) the link between the witness testimony and the fact that John shoots Peter (see Figure 5(a)) is attacked by questioning the truthfulness of the witness. Decompressed links may be compressed again in order to make the graph better readable by collapsing the expanded node. An attacked link that is decompressed is shown in Figure 5(c). Note that both causal and evidential links may be decompressed and may be subject to support or attack.

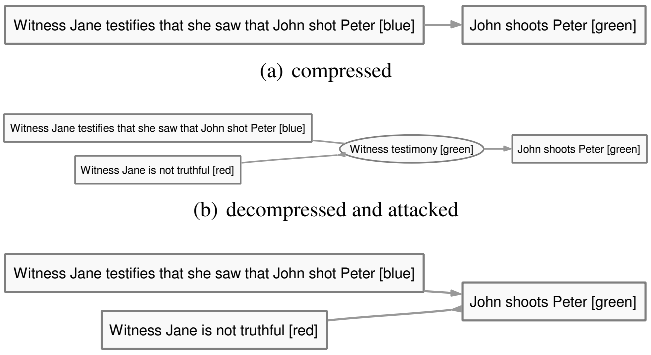

Peter is hit in the head [green]
Peter dies [green]
Peter is hit in the head [green]
Causal [yellow]
Peter dies [green]
Peter is hit in the head [green]
A man who is hit in the head by a bullet generally dies [green]
Causal [yellow]
Peter dies [green]

(a) compressed

(b) decompressed

(c) supported

Figure 4. Decompression and support of a causal link in *AVERs*


Witness Jane testifies that she saw that John shot Peter [blue]
John shoots Peter [green]
Witness Jane testifies that she saw that John shot Peter [blue]
Witness Jane is not truthful [red]
Witness testimony [green]
John shoots Peter [green]
Witness Jane testifies that she saw that John shot Peter [blue]
John shoots Peter [green]
Witness Jane is not truthful [red]

(a) compressed

(b) decompressed and attacked

(c) compressed after attack

Figure 5. Decompression and attack of an evidential link in *AVERs*

2 The study

The main purpose of this study was to test the effect of refinement and compression. The treatment group was allowed to use a system which contained the refinement and compression methods described above in order to analyze a simple crime case. The control group analyzed the same case by using a more basic system which contained a simpler method for collapsibility. This method allowed them to simultaneously collapse or expand all nodes of a certain type, while it was not possible to expand or collapse individual nodes. This functionality was offered to allow them to handle large stories on a basic level. In Figures 9 and 10 the difference between both conditions is displayed. Consider first Figure 8 in which the default view is shown where all links are compressed, that is, all inference nodes are hidden. The treatment group was able to decompress nodes and links one node at a time, so in Figure 9 the node `John shoots Peter` is expanded (it has a double border in order to show that it is expanded) and its surrounding links are decompressed. This is done by selecting the desired node and than clicking the "expand" link (see Figure 6). Expanded nodes may be collapsed again by clicking "collapse". The control group could only decompress the entire graph, that is, all nodes of a certain type may be hidden or displayed all at once, so in Figure 10 all links are expanded. Note that in Figure 9, only the causal links directly connected to the expanded node `John shoots Peter` are decompressed, while the others remain compressed. The advantage of the treatment group (and the difference between Figures 9 and 10) does not seem to be very important, but note that it will become more apparent when the size of the graph increases. In order to hide nodes of a certain type the users in the control group had to select the box next to it in a menu, as displayed in Figure 7 (note that only the first element of the menu is displayed, which allows users to decompress links, for the full menu see Figure 1). Unchecking this box again will display all nodes of this type. Contrary to the treatment group these users were not able to "zoom in" on a specific node in the graph and explore its status, because if they wanted to decompress a certain link they had to unfold all links. They were also not able to insert nodes by using the "insert node" menu (see Figure 6). We predicted that the participants in the treatment group would perform better, regarding the quality of their analysis and their understanding of the case, than the participants in the control group.

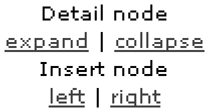

Detail node
expand | collapse
Insert node
left | right

Figure 6. Interface of decompression for the treatment group in *AVERs*

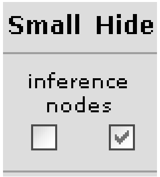

Small Hide
inference nodes
□ ☑

Figure 7. Interface of decompression for the control group in *AVERs*

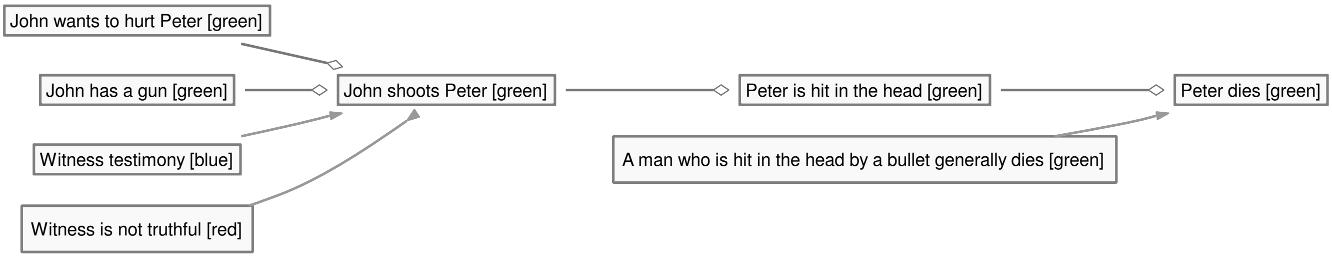

John wants to hurt Peter [green]
John has a gun [green]
Witness testimony [blue]
Witness is not truthful [red]
John shoots Peter [green]
Peter is hit in the head [green]
Peter dies [green]
A man who is hit in the head by a bullet generally dies [green]

Figure 8. A compressed story in *AVERs*

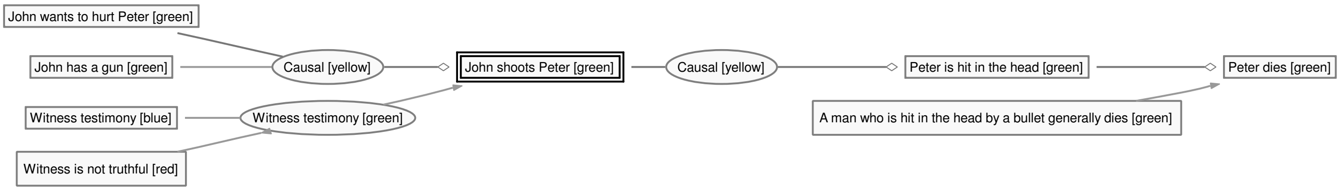

John wants to hurt Peter [green]
John has a gun [green]
Witness testimony [blue]
Witness is not truthful [red]
Causal [yellow]
Witness testimony [green]
John shoots Peter [green]
Causal [yellow]
Peter is hit in the head [green]
Peter dies [green]
A man who is hit in the head by a bullet generally dies [green]

Figure 9. Decompression for the treatment group in *AVERs*

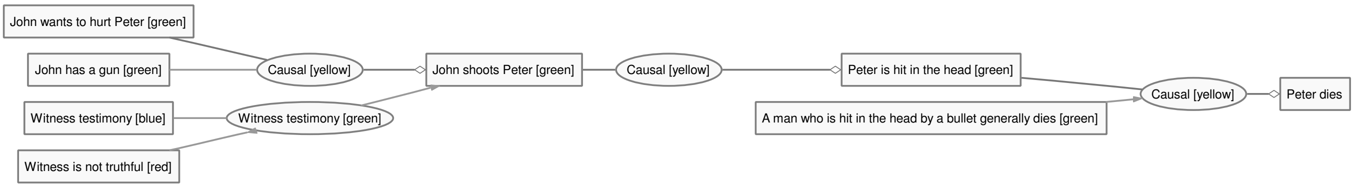

John wants to hurt Peter [green]
John has a gun [green]
Witness testimony [blue]
Witness is not truthful [red]
Causal [yellow]
Witness testimony [green]
John shoots Peter [green]
Causal [yellow]
Peter is hit in the head [green]
A man who is hit in the head by a bullet generally dies [green]
Causal [yellow]
Peter dies

Figure 10. Decompression for the control group in *AVERs*

2.1 Participants

The study was conducted during a three-hour session at the Dutch police academy in Zutphen; five crime analysis students and twelve analysts working in different districts in the Netherlands participated. However, one participant failed to complete the questionnaire and was excluded from the results ($N = 16$). Participants were assigned to the conditions randomly ($N = 8$ for both groups). To help to account for biases between the treatment and the control group, the participants' educational level, computer skills, and experience in conducting crime analyses and visualizing information were assessed by means of a questionnaire (pre-test).

2.2 Materials and procedure

In order to measure the effect of compression and refinement on the quality of the participants' analysis, a special questionnaire was devised that consisted of 74 questions or tasks. This questionnaire contained a pre-test, an interactive manual, the actual test, a post-test, and a usability questionnaire. The pre-test included several questions on the participants' skills and background. This test was mainly used to determine the population characteristics and to reveal pre-existing differences in education, computer skills, and experience in conducting crime analyses and visualizing information. The manual was only used to familiarize the participants with the system and its interface. In this part the participants were asked to visualize a sample story. While doing that, they were presented with information about the underlying concepts and with instructions on how to reproduce a certain part of the example. In the actual test the participants had to analyze a simplified murder case (note that due to time constraints the case that had to be analyzed was rather small and simple). They were provided with several source documents of evidence and were asked to use the software to analyze the case and construct graphs that represent their stories of what might have happened. This case consisted of five source documents (two witness statements, a lab report, a report of a house search, and a statement of a police officer) from which quotes could be selected in order to support events. To give an estimate of the size of the case, a completely constructed graph would contain 32 evidence and event nodes. The participants were allowed to use all functionality the system provides and to use the digital versions of the source documents that were available to them. After they handed in these first parts of the questionnaire they were asked to complete the post-test and the usability questionnaire. The post-test consisted of 16 true or false statements to test the participants' understanding of the case; of these statements eight tested the knowledge of important aspects of the case, while the other eight concerned minor details. When answering these questions, the participants were not allowed to read the evidential documents, but were permitted to use the stories they constructed earlier. Finally, the usability questionnaire contained 5-point Likert scale statements to test the user-friendliness of the system as a whole and the ease of use of specific features in it.

2.3 Dependent measures

Data was captured using a combination of logging of the participants' actions and graphs, and different questionnaires. The quality of the analysis was measured by assessing the quality of the produced graph. Graphs can be correct or wrong in different ways, for instance, a graph can be complete, that is, it contains all information present in the case. But a graph can also be well-structured or its containing argument can be sound. Therefore, the graphs were evaluated based on three criteria, namely completeness, structure, and soundness. Additionally, the time taken to produce the task was measured. T-Tests were conducted to evaluate the hypothesis that participants in the treatment group (who were allowed to use all compression and refinement methods) would perform better on their analysis of the sample case (higher quality stories) and understand the case better than those that did not use these methods. Note that all p-values that are reported are based on one-sided testing, that is, since we had directional predictions we tested our hypotheses with one-sided t-tests [9].

2.3.1 *Pre-test*

Pre-existing differences between the two groups were measured by three questions on the participants' experience in working with computers, visualizing information, and conducting crime analyses. On a scale from 0 to 4 the participants had to select whether they had no, little, average, or much experience in the particular domain ($MAX = 4$).

2.3.2 *Quality of the analysis*

The quality of the analysis is measured by scoring the quality of the produced graph by three indicators as assessed by an expert (the first author):

1. the completeness of the graph (for every correct element in the constructed graph the participant received 2 points, $MAX = 32$);
2. the number of correct links between the elements (structure; 1 point for every link in the correct direction, $MAX = 31$);
3. the soundness of the graph. The soundness was measured by the following criteria ($MAX = 20$)
   (a) the participant received 5 points if he correctly used arguments to attach evidence to stories;
   (b) he received 10 points if he correctly distinguished between causal and evidential links (5 points if he sometimes correctly used causal links, but in other cases confused them with evidential links);
   (c) the participant received 5 points if he correctly expressed reasons of doubt.

Subsequently, to obtain a global measure for quality these three indicators were summed. This measure was controlled for the range of its constituents by dividing the three indicators by their maximum score. This means that $overall = (completeness/32) + (structure/31) + (soundness/20)$ and that $MAX = 3$. Note that the assessment of graphs was done blind, such that the expert did not know whether the producer of the graph was in the treatment or the control group.

2.3.3 *Time taken for analysis*

The time taken was measured by the number of seconds that elapsed between the creation of the first node of the analysis and the last action taken on the case before logout.

2.3.4 *Understanding of the case*

The participants' understanding of the case was measured by the number of correct answers on 16 true or false statements about the analyzed case which were asked after the participants completed their analysis (post-test). These statements included 8 statements on important facts and 8 on smaller details of the case ($N = 15$, because one participant failed to complete the questionnaire). The participants received 1 point for every correct answer ($MAX = 8$ for the major and minor facts, and $MAX = 16$ for the total understanding).

2.3.5 Usability measures

The usability of the interface was measured by asking the participants to rate 14 statements on a 5-point scale, nine focused on the user-friendliness of the interface as a whole and five on the ease of use of specific features. The ease of use of the collapsibility feature was also measured separately. All measures are calculated by taking the sum of all ratings on the questions and dividing them by the number of questions ($MIN = 1$ and $MAX = 5$ for all measures).

3 Results

3.1 Pre-test scores

Pre-test scores revealed that there were no significant pre-existing differences between groups (see Table 1). In the remainder of this paper we will report t-test scores only, as controlling for pre-test scores is not necessary.

Table 1. Mean scores on pre-test for both conditions.

Measure	Treatment	Control	p
computer	2.38 ($SD = 0.52$)	2.25 ($SD = 0.46$)	.31
visualization	2.00 ($SD = 0.76$)	2.00 ($SD = 0.54$)	.50
crime analysis	2.13 ($SD = 0.84$)	2.00 ($SD = 0.76$)	.38

3.2 Effect of compression and refinement

3.2.1 Quality of the analysis

The mean scores of the treatment group were higher than those of the control group (see Table 2). A t-test showed that the difference in soundness was significant ($p = .04$). No other differences were statistically significant ($p = .40$ for completeness and $p = .24$ for structure), although these differences were in the expected direction. In total, the graphs of the treatment group were better than the graphs of the control group ($M = 1.01$ and $M = 0.73$ respectively). This difference was weakly significant ($p = .05$). On the whole the data suggest that the treatment groups produces higher quality analyses than the control group. We are aware of the fact that not all individual comparisons showed significant differences given the limited number of participants, yet the results show a pattern of differences in the predicted direction, therefore we decided to report and discuss all individual means.

Table 2. Mean scores on quality measures for both conditions.

Measure	Treatment	Control	p
completeness	16.25 ($SD = 5.90$)	15.50 ($SD = 5.63$)	.40
structure	6.00 ($SD = 2.83$)	4.75 ($SD = 3.88$)	.24
soundness	6.25 ($SD = 5.83$)	1.88 ($SD = 2.59$)	.04
overall	1.01 ($SD = 0.34$)	0.73 ($SD = 0.31$)	.05

3.2.2 Time taken for analysis

The participants in the control group ($M = 3972.13$ seconds with $SD = 876.91$) used more time than the participants in the treatment group ($M = 3285.88$ seconds with $SD = 724.70$). This difference was weakly significant $p = .06$.

3.2.3 Understanding of the case

Although the differences were not significant ($p > .05$), the participants in the treatment group performed somewhat better on the true or false statements than the control group (see Table 3). Furthermore, in the treatment group four participants were able to answer all questions correctly, while in the control group none of the participants was able to do so, showing that the treatment group developed a better understanding of the case during the analysis than the control group.

Table 3. Mean scores on understanding of the case (post-test) for both conditions.

Measure	Treatment	Control	p
major	7.25 ($SD = 0.89$)	7.00 ($SD = 0.58$)	.27
minor	7.63 ($SD = 0.74$)	7.14 ($SD = 1.07$)	.16
total	14.88 ($SD = 1.55$)	14.14 ($SD = 1.07$)	.16

3.3 Usability

3.3.1 User-friendliness

The first nine statements of the usability questionnaire measured the user-friendliness of the system as a whole, these questions were common to all subjects. For every participant all scores on the statements are summed and divided by 9 to obtain a measure of user-friendliness on a 5-point scale. It was found that $M = 2.80$ with $SD = 0.60$. This result means that this aspect needs improvement as a satisfactory score should be at least higher than 3 on a scale from 1 to 5.

3.3.2 Ease of use

The second five statements of the usability questionnaire measured the ease of use of specific features in *AVERs*, including the way nodes, links, quotes from source documents, and schemes are added to the graph. The questionnaire revealed a mean rating of 3.31 (with $SD = 0.60$) on a 5-point scale. More specifically, the participants in the treatment group ($M = 3.63$ with $SD = 0.92$) found the collapsibility feature easier to use than the participants in the control group ($M = 3.38$ with $SD = 1.06$), but this difference is not significant ($p = .31$). Nonetheless these results are satisfactory, as the scores were higher than 3 on a scale of 1 to 5.

4 Conclusion

The study suggest that crime analysts who are allowed to use methods to refine or compress links produce somewhat higher quality analyses and have a somewhat better understanding of the case than analysts who are provided with simpler methods to handle large graphs. All differences found between groups were in the expected direction, but only the difference in soundness was statistically significant, while the differences in the overall quality of the produced graph and the time taken to complete the task showed a trend in the predicted direction. On the whole the analyses presented in this paper indicate that the selected methods increase performance and they have shown the importance of suitable ways to handle large and complex graphs. The usability measures revealed that the ease of use of the features in *AVERs* is satisfactory but that the user-friendliness of the system as a whole needs improvement. In particular, with respect to user-friendliness the slowness of the system and the inability to undo actions were pointed out as drawbacks. While devising future versions of the system we will pay extra attention to these areas.

5 Discussion

Although the results presented in this paper are promising and in the predicted direction, the effects were not strong. Two reasons for this may be identified. Due to time constraints the case that had to be analyzed was rather small and simple. Arguably, in larger cases the differences between conditions might be even more apparent. However, we are aware of the fact that the usability of the interface needs to be improved before conducting new experiments, such that this does not influence the results. Additionally, we expect that repetition of this study with a larger number of participants will yield more significant results. Nonetheless, this study provides preliminary support for the claim that compression and refinement indeed support users better than conventional methods to produce readable graphs, while they also satisfy the specific needs of crime analysts. The insights presented here may be applied to similar argument visualization tools, including Araucaria [11], ATHENA [12], Belvedere [14], and Rationale [18], which lack these specific features. To our knowledge until now none of the conducted experiments on the effectiveness of these tools has proven any significant effect [15]. We expect that further studies will show significant effects for these tools if the refinement and compression methods we proposed are implemented. Thus our study can make a valuable contribution to visualization approaches to (evidential) reasoning.

In this study alternative solutions to handle large arguments, such as representation formats that do not use box and arrows were not included, although the effectiveness of matrix representations as an alternative to graphs has already been proven [13, 5]. Therefore, in future studies we will test the effect of these alternative representation formats on the users' performance.

ACKNOWLEDGEMENTS

This research was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 634.000.429. The authors would like to thank the Dutch police academy in Zutphen and particularly Gerard Bijsterbosch and Henk Evers for their cooperation.

REFERENCES

[1] i2 Analyst's Notebook: Investigative analysis software (www page). `http://www.i2.co.uk/Products/Analysts_Notebook/default.asp`, 2006.

[2] Floris J. Bex, Henry Prakken, Chris A. Reed, and Douglas N. Walton, 'Towards a formal account of reasoning about evidence: Argumentation schemes and generalisations', *Artificial Intelligence and Law*, **11**, 125–165, (2003).

[3] Floris J. Bex, Susan W. van den Braak, Herre van Oostendorp, Henry Prakken, Bart Verheij, and Gerard A.W. Vreeswijk, 'Sense-making software for crime investigation: How to combine stories and arguments?', *Law, Probability and Risk*, **6**(1–4), 145–168, (2007).

[4] C.J. de Poot, R.J. Bokhorst, Peter J. van Koppen, and E.R. Muller, *Rechercheportret: Over Dilemma's in de Opsporing (Investigation Portrait: About Dilemmas in Investigation)*, Kluwer, Den Haag, The Netherlands, 2004.

[5] Mohammad Ghoniem, Jean-Daniel Fekete, and Philippe Castagliola, 'On the readability of graphs using node-link and matrix-based representations: A controlled experiment and statistical analysis', *Information Visualization*, **4**(2), 114–135, (2005).

[6] Paul A. Kirschner, Simon J. Buckingham Shum, and Chad S. Carr, *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, Springer-Verlag, London, UK, 2003.

[7] Ron P. Loui and Jeff Norman, 'Rationales and argument moves', *Artificial Intelligence and Law*, **3**(3), 159–189, (1995).

[8] John D. Lowrance, 'Graphical manipulation of evidence in structured arguments', *Law, Probability and Risk*, **6**(1–4), 225–240, (2007).

[9] Robert B. McCall, *Fundamental Statistics for Psychology*, Harcourt Brace Jovanovich, Inc, New York, NY, 1975.

[10] Nancy Pennington and Reid Hastie, *Inside the Juror: The Psychology of Juror Decision Making*, chapter The story model for juror decision making, Cambridge University Press, 1993.

[11] Chris A. Reed and Glenn W.A. Rowe, 'Araucaria: Software for argument analysis, diagramming and representation', *International Journal of Artificial Intelligence Tools*, **14**(3-4), 961–980, (2004).

[12] Bertil Rolf and Charlotte Magnusson, 'Developing the art of argumentation: A software approach', in *Presented at the Fifth International Conference on Argumentation*, (2002).

[13] Daniel D. Suthers and Christopher D. Hundhausen, 'An empirical study of the effects of representational guidance on collaborative learning', *Journal of the Learning Sciences*, **12**(2), 183–219, (2003).

[14] Daniel D. Suthers, Arlene Weiner, John Connelly, and Massimo Paolucci, 'Belvedere: Engaging students in critical discussion of science and public policy issues', in *AI-Ed 95, The 7th World Conference on Artificial Intelligence in Education*, pp. 266–273, (1995).

[15] Susan W. van den Braak, Herre van Oostendorp, Henry Prakken, and Gerard A.W. Vreeswijk, 'A critical review of argument visualization tools: Do users become better reasoners?', in *Workshop Notes of the ECAI-2006 Workshop on Computational Models of Natural Argument (CMNA VI)*, eds., Floriana Grasso, Rodger Kibble, and Chris Reed, pp. 67–75, (2006).

[16] Susan W. van den Braak and Gerard A.W. Vreeswijk, 'AVER: Argument visualization for evidential reasoning', in *Legal Knowledge and Information Systems. JURIX 2006: The Nineteenth Annual Conference*, ed., Tom M. van Engers, pp. 151–156, Amsterdam, The Netherlands, (2006). IOS Press.

[17] Susan W. van den Braak, Gerard A.W. Vreeswijk, and Henry Prakken, 'AVERs: An argument visualization tool for representing stories about evidence', in *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, pp. 11–15, New York, NY, (2007). ACM Press.

[18] Tim J. van Gelder, 'The rationale for Rationale^{TM}', *Law, Probability and Risk*, **6**(1–4), 23–42, (2007).

[19] Bart Verheij, *Virtual Arguments. On the Design of Argument Assistants for Lawyers and Other Arguers*, TMC Asser Press, The Hague, The Netherlands, 2005.

[20] Willem A. Wagenaar, Hans F.M. Crombag, and Peter J. van Koppen, *Anchored Narratives: Psychology of Proof in Criminal Law*, St Martin's Press / Prentice-Hall, New York, NY, 1993.

[21] Douglas N. Walton, *Argumentation Schemes for Presumptive Reasoning*, Lawrence Erlbaum Associates, Mahwah, NJ, 1996.