The role of compression in visualization tools for crime analysts

Susan W. van den Braak, Herre van Oostendorp, Gerard A.W. Vreeswijk, Department of Information and Computing Sciences, Utrecht University, PO Box 80.089, 3508 TB Utrecht, The Netherlands
Email: susanb@cs.uu.nl, herre@cs.uu.nl, gv@cs.uu.nl
Henry Prakken, Department of Information and Computing Sciences, Utrecht University & Faculty of Law, University of Groningen, PO Box 716, 9700 AS Groningen, The Netherlands, henry@cs.uu.nl

Abstract:

Introduction

Graph visualization tools have been shown promising for learning and for collaborating on the construction and evaluation of arguments (Kirschner, Buckingham Shum, & Carr, 2003; van den Braak, van Oostendorp, Prakken, & Vreeswijk, 2006). Increasingly, such tools are applied to the domain of reasoning about evidence in legal settings (Verheij, 2005; Lowrance, 2007; van Gelder, 2007). 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 the same spirit, a tool for the visualization of stories and evidence named *AVERs* (Argument Visualization for Evidential Reasoning based on stories) has been developed (van den Braak, Vreeswijk, & Prakken, 2007; Bex, van den Braak, van Oostendorp, Prakken, Verheij, & Vreeswijk, 2007). This tool combines such sense-making tools, which focus on argumentation, with a story-based approach from legal psychology called Anchored Narratives Theory (Wagenaar, van Koppen, & Crombag, 1993).

In the current practice of crime analysis, software for managing and visualizing evidence is already being used, that allow them to formulate stories as simple timelines. A well-known example is Analyst's Notebook (http://www.i2.co.uk/Products/AnalystsNotebook/default.asp). 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; the can construct possible stories about what happened by linking events through causal connections and can link the evidence that they find 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 to structure the information they gather (de Poot, Bokhorst, van Koppen, & Muller, 2004). 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.

Although such a visualization approach seems promising, and is often claimed to be beneficial and to provide faster learning, 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 (van den Braak, van Oostendorp, Prakken, & Vreeswijk, 2006). 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. It will become more difficult for them to read and interact with these graphs as they have to keep overview of the complete graph, keep track of changes, and quickly finding 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 inverting 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 displayed 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. 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 the smallest element of a story, explore its status, and explain it in more detail 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, 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 compression methods, based on the compression rationale (Loui & Norman, 1995), are a viable way to provide these layers, while they also improve the readability of graphs. Compression 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". To attack this rule, we have to restate the argument 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 can be useful for arguments that are not subject to attack and maintains the graph orderly. However, sometime rules need to be decompressed in order to make the underlying reasoning explicit. Useful software tools for crime analysts should provide two compression methods, that is, features to refine links and functionality to fold or unfold links in order to add reasons. For these are both tasks that are necessary and important during the analysis of a case.

Although such compression methods seem promising and useful, most state-of-the-art sense-making tools, such as Araucaria (Reed & Rowe, 2004) and Rationale (Rolf & Magnusson, 2002), only allow for zooming in and out on large graphs and do not provide more advanced compression methods. Besides, little research is done on the effect of compression on the users' performance. This paper reports on a study that was conducted to determine whether tools that contain such compression methods support their users better than tools that only provide standard methods to handle large graphs, for example zooming features. We have done this by measuring the effect of compression on the quality of crime analyses.

The *AVERs* system

The *AVERs* software can be used to draw 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. 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 colour pictures has been put online at [url].

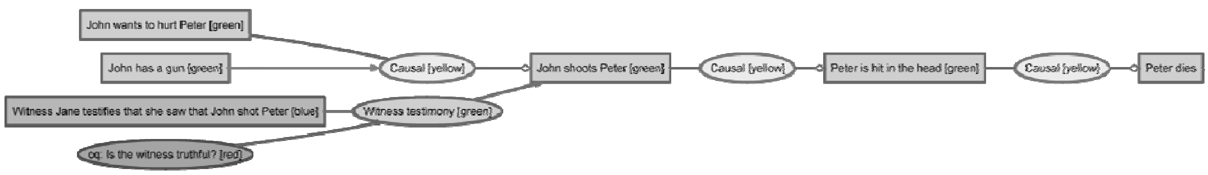

Figure 1. A simple story in *AVERs*.

In Figure 1 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. Take for example the simple story depicted, the observation that "John wants to hurt Peter" causes the fact that he shoots him. 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. Typically, 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 argumentation 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 a 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 not answered yet). Scheme instantiations are represented as ellipses that display the type of the scheme that was used. 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 (van den Braak & Vreeswijk, AVER: Argument visualization for evidential reasoning, 2006; van den Braak, Vreeswijk, & Prakken, 2007).

Compression methods

We incorporated two compression methods in *AVERs*, namely refinement and collapsibility. Refinement allows for the addition of a new node in between two previously connected nodes to refine stories. Consider for example Figure 2, suppose that at first the analyst did create 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).

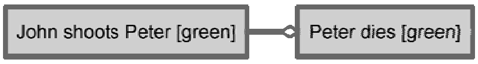

Figure 2. A causal link in *AVERs*.

Figure 3. A refined causal link in *AVERs*.

Secondly, collapsibility involves the folding and unfolding of redundant nodes and links. By default all links are collapsed, but users may expand these nodes if they want to add defeaters to attack them. In this way decompression points out possibilities for attack and allows for underlying reasons to be made explicit. In Figure 5 the link between "John wants to hurt Peter" and "John shoots Peter" from Figure 4 was expanded to add a reason for this link, namely that the fact that John owns a gun caused him to shoot Peter. Similarly, collapsibility may be used to attack links; this is what is done automatically when critical questions are answered negatively.

Figure 4. A collapsed causal link in *AVERS*.

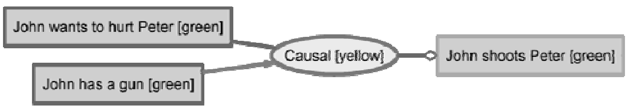

Figure 5. An expanded and supported causal link in *AVERs*.

The study

Subjects

The study was conducted during a three hour session at the Dutch police academy in Zutphen; 5 crime analysis students and 12 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). Subjects 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 subjects' educational level, computer skills, and experience in conducting crime analyses and visualizing information were assessed by means of a questionnaire (pre-test).

The treatment group was allowed to use a system which contained the two methods for compression, namely refinement and collapsibility (on both node and graph level) in order to analyze a simple crime case. The control group analyzed the same case by using a more basic system which contained only a simple method for collapsibility. This method allowed them to 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. We predicted that the subjects in the treatment group would perform better than the subjects in the control group.

Materials and procedure

In order to measure the effect of compression on the quality of the subjects' analysis, a special questionnaire containing five parts was devised. The first part included several questions on the subjects' skills and background. This pre-test was mainly used to determine the population characteristics and pre-existing differences in education, computer skills, and experience in conducting crime analyses and visualizing information. The second part was only used to familiarize the subjects with the system and it's interface. In this part the subjects were asked to reproduce a sample story. While doing that they were presented with information about the underlying concepts and instructions on how to reproduce a certain part of the example.

In the actual test, Part C, the subjects had to analyze a small, simplified murder case. They were provided with several source documents of evidence such as testimonies and were asked to use the software to analyze the case and construct graphs that represent their stories of what might have happened. The subjects were allowed to use all functionality the system provides and to use the digital versions of the software that were available to them.

After they handed in the first parts of the questionnaire they were asked to complete the last two parts of the questionnaire. Part D (post-test) consisted of 16 true or false statements to test the subjects' understanding of the case; of these statements 8 tested the knowledge of important aspects of the case, while the other 8 concerned minor details. When answering these questions, the subjects were not allowed to read the evidential documents, but were permitted to use the stories they constructed earlier. Finally, the last part contained 5-point Likert scale statements to test the general user-friendliness and ease of use of the system.

Dependent measures

Data was captured using a combination of logging of the subjects' actions and graphs, and different questionnaires. The quality of the analysis was measured by expert assessment. The produced 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. Additionally, the time taken to produce the task is measured. T-Tests were conducted to evaluate the hypothesis that subjects in the treatment group (who were allowed to use all compression 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 compression. Note that all p-values that are reported are based on 1-sided testing.

Pre-test

Pre-existing differences between the two groups were measured by three questions on the subjects' experience in working with computers, visualizing information, and conducting crime analyses. On a scale from 0 to 4 the subjects had to select whether they had none, little, average, or much experience in the particular domain (MAX = 4).

Quality of analysis

This variable was measured by three indicators as assessed by an expert (the author):

1. the completeness of the graph (*complete*, for every correct element in the constructed graph the subject 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 = 16);
3. the soundness of the graph (*sound*). The soundness was measured by the following criteria (MAX = 20)
    a. the subject 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); and
    c. the subject received 5 points if he correctly expressed reasons of doubt.

Subsequently, controlling for the range of its constituents, these three indicators are divided by their maximum score and summed in order to get a total, but relative, measure of the quality called *overall* (MAX = 3). This means that *overall* = (*complete* / 32) + (*structure* / 16) + (*sound* / 20).

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.

Understanding of case

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

Usability measures

The usability of the interface was measured by asking the subjects to rate 14 statements on a 5-point scale, 9 focussed on the usability of the interface (*use*) and 5 on its ease of use (*ease*). The ease of use of the collapsibility feature was also measured separately through *collapse* (also included in *ease*). 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).

Results

Pre-test scores

Pre-test scores revealed that there were no significant pre-existing differences between groups. 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

Effect of compression methods

Quality of analysis

The graphs that were produced by the treatment group were more complete, better structured and more sound than the graphs produced by 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 those differences were in the expected direction. In total, the graphs of the treatment group were better than the graphs of the control group (M=1.22 and M=0.89 respectively). This difference was marginally significant (p=.07). On the whole the data suggests that the compression users produce higher quality analyses than the control group.

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

	complete	*structure*	*sound*	*overall*
Treatment	16.25 (SD = 5.90)	6.00 (SD = 2.83)	6.25 (SD = 5.83)	1.22 (SD = 0.40)
Control	15.50 (SD = 5.63)	4.75 (SD = 3.88)	1.88 (SD = 2.59)	0.89 (SD = 0.41)

Time taken for analysis

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

Understanding of case

The subjects in the treatment group performed better on the true or false statements than the control group (see Table 3). In the treatment group 4 subjects were able to answer all questions correctly, in the treatment group none of the subjects was able to do so. However, the differences were not significant.

Table 3: Mean scores on 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
understanding	14.88 (SD = 1.55)	14.14 (SD = 1.07)	.16

Usability

User-friendliness

The first nine statements of Part E of the questionnaire measure the user-friendliness of the software. For every subject all scores on the statements are summed and divided by 9 to obtain a measure of user-friendliness. It was found that M = 2.80 with SD = 0.60. This aspect needs improvement as a satisfactory score should be at least higher than 3 on a scale from 1 to 5.

Ease of use

The second five statements of Part E measure the ease of use of the software. It was found that M = 3.31 with SD = 0.60. The subjects in the treatment group (M = 3.63 with SD = 0.92) found the collapsibility feature easier to use than the subjects in the control group (M = 3.38 with SD = 1.06), but this difference is not significant (p = .31). These results are satisfactory, as the scores were higher than 3 on a scale of 1 to 5.

Conclusion

The study showed that crime analysts who are allowed to use compression methods to refine or expand links produce higher quality analyses 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 were nearly significant and showed a trend in the predicted direction. On the whole the analysis presented in this paper indicates that the selected methods increase performance and it has shown the importance of suitable ways to handle large and complex graphs. The usability measures revealed that the user-friendliness and ease of use of AVERs are satisfactory but need improvement. In particular, 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.

Discussion

It should be noted that 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. Additionally, we expect that repetition of this study with a larger number of subjects will yield more significant results.

hier nog referen naar related work

Although the effectiveness of matrix representations as an alternative to graphs has already been proven (Suthers & Hundhausen, 2003; Ghoniem, Fekete, & Castagliola, 2005), in this study other solutions to work with large graphs, such as representation formats that do not use box and arrows were not included. Therefore, in future studies we will test the effect of these alternative representation formats on the users' performance.

References

Bex, F., van den Braak, S., van Oostendorp, H., Prakken, H., Verheij, B., & Vreeswijk, G. (2007). Sense-making software for crime investigation: how to combine stories and arguments? *Law, Probability and Risk Advance Access published on July 7, 2007* (doi:10.1093/lpr/mgm007).

de Poot, C. J., Bokhorst, R. J., van Koppen, P. J., & Muller, E. R. (2004). *Rechercheportret: Over Dilemma's in de Opsporing.* Den Haag, The Netherlands: Kluwer.

Ghoniem, M., Fekete, J.-D., & Castagliola, P. (2005). On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. *Information Visualization , 4* (2), 114-135.

Kirschner, P., Buckingham Shum, S. J., & Carr, C. S. (2003). *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making.* London, UK: Springer.

Loui, R. P., & Norman, J. (1995). Rationales and argument moves. *Artificial Intelligence and Law. 3,* pp. 159-189. Springer.

Lowrance, J. D. (2007). Graphical manipulation of evidence in structured arguments. *Law, Probability and Risk Advance Access published on September 25, 2007* (doi:10.1093/lpr/mgm011).

Reed, C. A., & Rowe, G. W. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal of Artificial Intelligence Tools , 14* (3-4), 961-980.
Rolf, B., & Magnusson, C. (2002). Developing the art of argumentation: A software approach. *Proceedings of the Fifth Conference of the International Society for the Study of Argumentation.* Amsterdam.
Suthers, D. D., & Hundhausen, C. D. (2003). An Empirical Study of the Effects of Representational Guidance on Collaborative Learning. *Journal of the Learning Sciences , 12* (2), 183-219.
van den Braak, S. W., & Vreeswijk, G. A. (2006). AVER: Argument visualization for evidential reasoning. In T. M. van Engers (Red.), *Legal Knowledge and Information Systems. JURIX 2006: The Nineteenth Annual Conference* (pp. 151-156). Amsterdam, The Netherlands: IOS Press.
van den Braak, S. W., van Oostendorp, H., Prakken, H., & Vreeswijk, G. A. (2006). A critical review of argument visualization tools: Do users become better reasoners? In F. Grasso, R. Kibble, & R. A. Chris (Red.), *Workshop Notes of the ECAI-2006 Workshop on Computational Models of Natural Argument*, (pp. 67-75). Riva del Garda, Italy.
van den Braak, S. W., Vreeswijk, G. A., & Prakken, H. (2007). AVERs: an argument visualization tool for representing stories about evidence. *Proceedings of the 11th international conference on Artificial intelligence and law* (pp. 11-15). New York, NY: ACM Press.
van Gelder, T. (2007). The rationale for RationaleTM. *Law, Probability and Risk Advance Access published on October 10, 2007* (doi:10.1093/lpr/mgm032).
Verheij, B. (2005). *Virtual Arguments. On the Design of Argument Assistants for Lawyers and Other Arguers.* Den Haag: T.M.C. Asser Press.
Wagenaar, W. A., van Koppen, P. J., & Crombag, H. F. (1993). *Anchored Narratives. The Psychology of Criminal Evidence.* New York, NY: St Martins Press.

Acknowledgments

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.