Data Coherence Diagnosis in BBN Risky Behavior Model

Aleksandra V. Toropova

Abstract A problem of coherence diagnosis for risky behavior model based on the data about behavior episodes retrieved from an interview with a respondent is considered. The extension of the model is described and the examples of data coherency diagnostics are provided. For more convenient work with suggested method the software is provided.

Keywords Bayesian belief network · BBN · Data coherence diagnosis

1 Introduction

Many studies on Artificial Intelligence face with the problem of data coherence diagnosis [1–4]. Many models and decision-making systems include expert knowledge and expert estimates. Expert methods are used in situations where selection, justification and impact assessment cannot be made using exact calculations [5]. Researchers can use the information received from the experts only if they have a possibility to present it in a form suitable for further research. Therefore, it is necessary to either formalize the experts' knowledge, or evaluate its coherence and reliability [6, 7]. Such problems arise in studies focused on troubleshooting power system [4, 8], the diagnosis of distributed systems security problems anomalies [1, 9], as well as in other areas. In all these cases, the coherence of the data is extremely important.

In many fields of sociological, psychological and marketing research, we face the problem of risky behavior rate or frequency estimate on the basis of respondents' self-reports about their behavior. We need to estimate behavior rate using the responses to the questionnaire or the results of the interview [8, 10].

A.V. Toropova (✉)
SPIIRAS, 39, 14 Line, Saint Petersburg 199178, Russia
e-mail: alexandra.toropova@gmail.com

A. Abraham et al. (eds.), *Proceedings of the First International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'16)*, Advances in Intelligent Systems and Computing 451, DOI 10.1007/978-3-319-33816-3_12

An approach to the risky behavior rate estimate based on Bayesian belief networks and data obtained from interviews about last episodes of respondent's behavior is proposed in [10–12].

The initial model [11] was based on the data about the three latest episodes of respondents' risky behavior and minimum and maximum intervals between the episodes. These data were usually obtained from questionnaires or interviews [13]. Respondents could give false (not corresponding to actual behavior) answers to make a positive impression or due to memory-related issues: episodes of risky behavior could happen a long time ago and, hence, be hard to remember. For example, the risky sexual behavior data (information needed for decision-making in various areas, including education, medicine and public health) was under-reported very often due to its very private nature, and such data often became a subject of significant social desirability bias. Sometimes respondents answering question could make mistakes or be confused [14, 15].

As an example, consider the following scenario. During an interview conducted on Monday, the respondent replied that the last behavior episode was on the last Monday, the previous one was on the last Wednesday and the last but two episode was a month ago. At the same time, the respondent defined the minimum interval between episodes as a "week". Hence, the provided data were incoherent because the interval between the last episode and the previous one was less than the minimum.

Note, that this is a very simplified example of the problem; obviously such inconsistencies can be easily identified. However, there are possible more complex situations because of the sampling variables included in the model.

Thus, applications that used data obtained from respondents often faced with the problem of incoherent data. Therefore it is important to have tools to diagnose such situations.

In the paper we describe modified model that solves this problem. For more convenient work with the model software is provided. Also we discuss an extended example of the model usage.

2 Model Description

Figure 1 shows a generalized risky behavior model $M = (G(V, L), \mathbf{P})$ as a Bayesian belief network [16, 17]. The model structure is represented by the directed graph $(G(V, L))$, where $V = \{t_{01}, t_{12}, t_{23}, t_{\min}, t_{\max}, \lambda, n\}$ is corresponded to the set of nodes, $L = \{(u, v) : u, v \in V\}$ is corresponded to the set of directed links between nodes. In other words, Fig. 1 shows random elements included in the model and relations between them. We used GeNIe 2.0 [18] to create Bayesian belief network and to implement the probabilistic reasoning algorithms. All figures were also constructed in GeNIe 2.0.

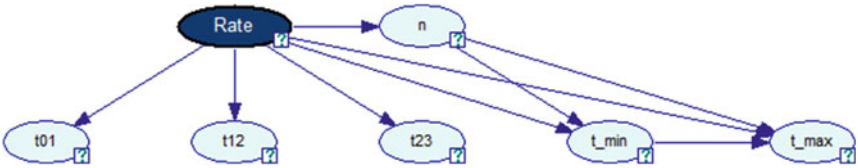

Fig. 1 Risky behavior model on the basis of data about episodes

On Fig. 1, *Rate* is a random variable representing the behavior rateλ, $t_{i,j}$ are random variables characterizing the lengths of the interval between the *i*th and *j*th to the end episodes. With an assumption that behavior was a Poisson random process random variables $t_{i,j}$ were exponentially distributed. The additional information was obtained by including minimum and maximum intervals between episodes ($t_{\min}$ and $t_{\max}$ respectively).

We specified conditional probabilities $\mathbf{P} = \{P(t_{j,j+1}|\lambda), P(t_{01}|\lambda), P(t_{\min}|n,\lambda), P(t_{\max}|n,\lambda,t_{\min}), P(n|\lambda), P(\lambda)\}$, (edges between conditionally dependent nodes) as follows:

$$
\begin{aligned}
P\left(t_{j,j+1}^{l_i j}|\lambda^{(i)}\right) &= e^{-a\lambda^{(i)}} - e^{-b\lambda^{(i)}},\ j = 0,1,2,\\
t_{j,j+1}^{l_i j}[a;b]; P\left(t_{\min}^{l_3}|n,\lambda^{(i)}\right) &= e^{-an\lambda^{(i)}} - e^{-bn\lambda^{(i)}},\ t_{\min}^{l_3} = [a;b];\\
p\left(n\middle|\lambda^{(i)}\right) &= \frac{\left(\lambda^{(i)}T\right)^n}{n!}e^{-\lambda^{(i)}T};\\
p\left(t_{\max}^{(l_4)}\middle|n,\lambda^{(i)},t_{\min}^{(l_3)}\right) &= e^{(n-1)\lambda^{(i)}t_{\min}^{(l_3)}}\left(\left(e^{-\lambda^{(i)}t_{\min}^{(l_3)}} - e^{-\lambda^{(i)}b}\right)^{n-1} - \left(e^{-\lambda^{(i)}t_{\min}^{(l_3)}} - e^{-\lambda^{(i)}a}\right)^{n-1}\right),\\
t_{\max}^{(l_4)} &= [a;b).
\end{aligned}
$$

3 Model Extension

Figure 2 shows extended risky behavior model. The added nodes allow to estimate data given by a respondent.

The nodes $c_{t_{1,2,\min}}$ and $c_{t_{23,\min}}$ represent episode t_{ij} and minimal interval $t_{\min}$ coherence, the nodes $c_{t_{0,1,\max}}, c_{t_{12,\max}}$ and $c_{t_{23,\max}}$ represent episode t_{ij} and maximal interval $t_{\max}$ coherence. We did not consider $c_{t_{0,1,\min}}$, because t_{01} represents an interval between an risky behavior episode and the moment of interview, which is not an observing behavior episode.

In particular, for the node representing the coherence degree with a minimum interval, coherence rate $c_{t_{ij,\min}}$ could take the following three values: the values:t_{ij} and t_{min} were *coherent* ($c^{+}_{t_{ij,\min}}$), values were *incoherent* ($c^{-}_{t_{ij,\min}}$) and values were *undefined* ($c^{?}_{t_{ij,\min}}$). We assumed that the rate $c_{t_{ij,\min}}$ was undefined when both t_{ij} and

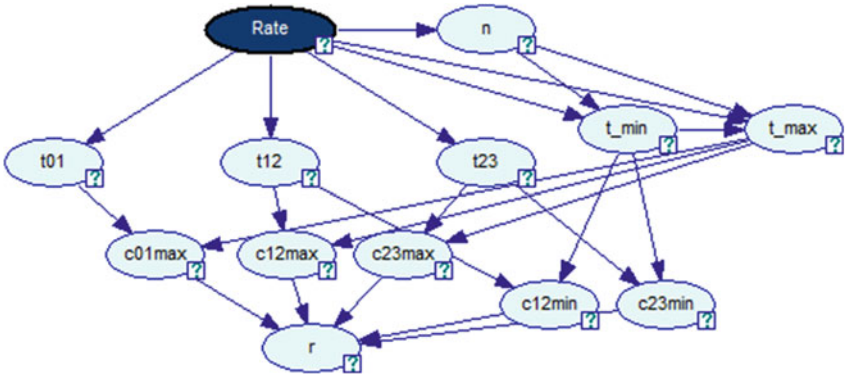

Fig. 2 Extended risky behavior model on the basis of data about episodes

t_{min} belong to the same intervals, i.e. if $t_{ij} \in [a;b)$ and $t_{min} \in [a;b)$ we could not define precisely whether the value t_{min} was smaller than t_{ij} or not.

We specified conditional probabilities of the extended model as follows:

$$P\left(c^{(s)}_{t_{ij,\min}} | t_{ij}, t_{\min}\right) = \begin{cases} \alpha^{(s)}, & t_{ij} > t_{min}; \\ \beta^{(s)}, & t_{ij} < t_{min}; \\ 1 - \alpha^{(s)} - \beta^{(s)}, & t_{ij} = t_{\min}; \end{cases}$$

where $s \in \{+, -, ?\}, \alpha^{(s)}, \beta^{(s)} \in [0;1], \sum \alpha = 1, \sum \beta = 1, \alpha^{(s)} + \beta^{(s)} \leq 1$.

Similarly, we obtained the estimation of the coherence of the random variables t_{ij} corresponding to the intervals between the last episodes realizations with the realization of a random variable $t_{\max}$ ($c_{t_{0,1,\max}}, c_{t_{12,\max}}$ and $c_{t_{23,\max}}$):

$$p\left(c^{(s)}_{t_{ij,\max}} | t_{ij}, t_{\max}\right) = \begin{cases} \alpha^{(s)}, & t_{ij} < t_{\max}; \\ \beta^{(s)}, & t_{ij} > t_{\max}; \\ 1 - \alpha^{(s)} - \beta^{(s)}, & t_{ij} = t_{\max}; \end{cases}$$

where $s \in \{+, -, ?\}, \alpha^{(s)}, \beta^{(s)} \in [0;1], \sum \alpha = 1, \sum \beta = 1, \alpha^{(s)} + \beta^{(s)} \leq 1$.

To estimate respondent reliability ($\boldsymbol{r}$) we added a node connecting all these five new nodes characterizing the pairwise coherence.

To simplify the formulae for conditional probabilities let $c = \left(c_{t_{12,\min}}, c_{t_{23,\min}}, c_{t_{01,\max}}, c_{t_{12,\max}}, c_{t_{23,\max}}\right)$, $c^{+} = \left(c^{+}_{t_{12,\min}}, c^{+}_{t_{23,\min}}, c^{+}_{t_{01,\max}}, c^{+}_{t_{12,\max}}, c^{+}_{t_{23,\max}}\right)$, $c^{-} = \left(c^{-}_{t_{12,\min}}, c^{-}_{t_{23,\min}}, c^{-}_{t_{01,\max}}, c^{-}_{t_{12,\max}}, c^{-}_{t_{23,\max}}\right)$, $c^{?} = \left(c^{?}_{t_{12,\min}}, c^{?}_{t_{23,\min}}, c^{?}_{t_{01,\max}}, c^{?}_{t_{12,\max}}, c^{?}_{t_{23,\max}}\right)$.

Then $p(\boldsymbol{r}^{+}|c) = \frac{\sum c^{+}}{\sum c}, p(\boldsymbol{r}^{-}|c) = \frac{\sum c^{-}}{\sum c}$ and $p\left(\boldsymbol{r}^{?}|c\right) = \frac{\sum c^{?}}{\sum c}$.

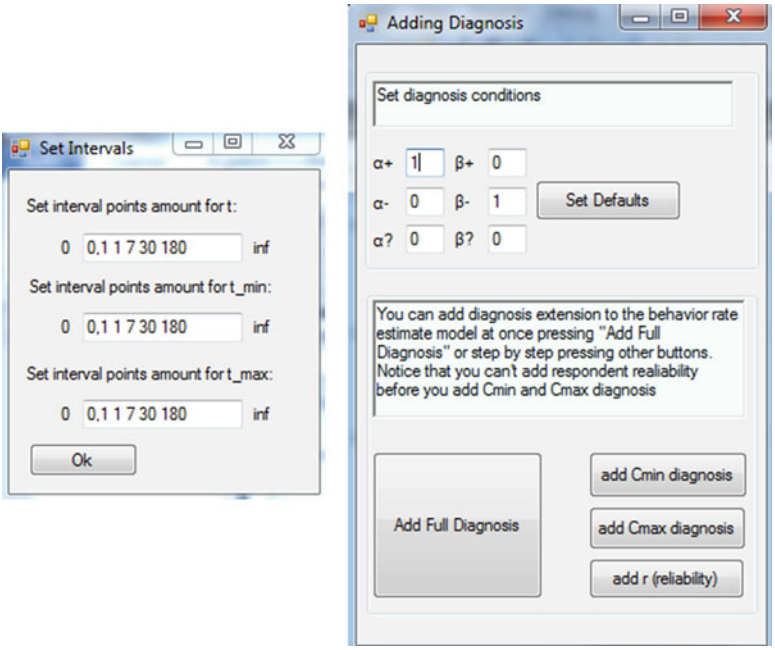

Fig. 3 Setting intervals and adding diagnosis windows

4 Realization

We created software to supplement the risky behavior model with mentioned before diagnosis nodes and for more convenient work with this model. The software was developed by using C# and Smile library [18]. Firstly user defines model: sets intervals for t_{ij}, t_{min} and $t_{\max}$ (Fig. 3); sets $\alpha^{(s)}, \beta^{(s)}$ where $s \in \{+, -, ?\}$ and add diagnosis nodes to the model, it can be made at once or step by step (Fig. 3). After that respondents data can be inserted into the model, input can be made manually then results are shown in the same window (Fig. 4) or from MS Excel file in this case results are saved in a separate file.

5 Example

Let t_{ij} to be divided into these disjunctive intervals: $t^{(1)} = (0; 0, 1)$, $t^{(2)} = [0, 1; 1)$ $t^{(3)} = [1; 7)$, $t^{(4)} = [7; 30)$, $t^{(5)} = [30; 180)$, $t^{(6)} = [180; +\infty)$, for clarity we take the same partition for $t_{\min}$ and $t_{\max}$.

We assumed that the coherence probability was zero, if the data provided by the respondent contradicted each other, and one, if there were no contradictions.

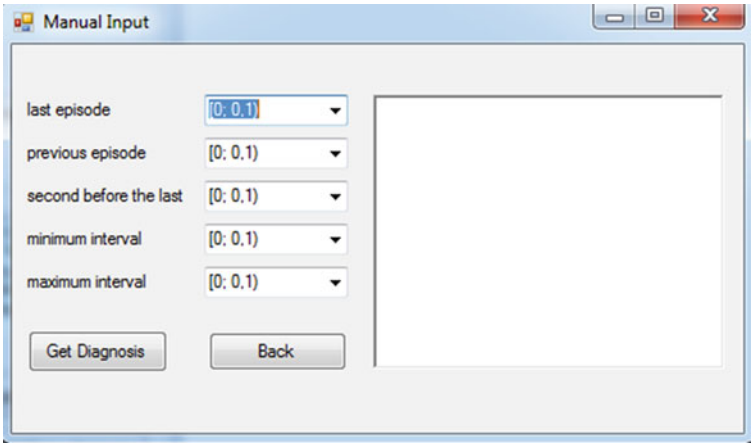

Fig. 4 Manual input window

We considered the example with ten respondents' given data. The data are presented in the Table 1, the first column contains a respondent's id, the other columns contain respondent's evidences about last risky behavior episodes (t_{01}, t_{12}, t_{23}) and minimal and maximal interval evidences ($t_{\min}$ and $t_{\max}$).

Let us have a closer look to the second respondent's data, particularly the coherence estimation of episode t_{12} and minimal interval $t_{\min}$. In this case data is incoherent. The posterior distribution of the coherence random variable is shown in Fig. 5.

After all the coherence estimations were defined, we estimated the respondent's reliability. The second respondent's reliability estimation is presented in Fig. 6.

Table 1 Respondents' data

Respondent's id	t_{01}	t_{12}	t_{23}	$t_{\min}$	$t_{\max}$
1	$[0, 1; 1)$	$[1; 7)$	$[0, 1; 1)$	$(0; 0, 1)$	$[7; 30)$
2	$[7; 30)$	$[0, 1; 1)$	$[7; 30)$	$[1; 7)$	$[30; 180)$
3	$[7; 30)$	$[1; 7)$	$(0; 0, 1)$	$(0; 0, 1)$	$[7; 30)$
4	$[0, 1; 1)$	$[0, 1; 1)$	$[0, 1; 1)$	$(0; 0, 1)$	$[180; +\infty)$
5	$[30; 180)$	$[7; 30)$	$[1; 7)$	$(0; 0, 1)$	$[180; +\infty)$
6	$[0, 1; 1)$	$[0, 1; 1)$	$[0, 1; 1)$	$(0; 0, 1)$	$[0, 1; 1)$
7	$[1; 7)$	$[1; 7)$	$[1; 7)$	$[0, 1; 1)$	$[7; 30)$
8	$[30; 180)$	$[30; 180)$	$[30; 180)$	$[30; 180)$	$[30; 180)$
9	$[180; +\infty)$	$(0; 0, 1)$	$[180; +\infty)$	$[0, 1; 1)$	$[30; 180)$
10	$[7; 30)$	$[7; 30)$	$[30; 180)$	$[1; 7)$	$[180; +\infty)$

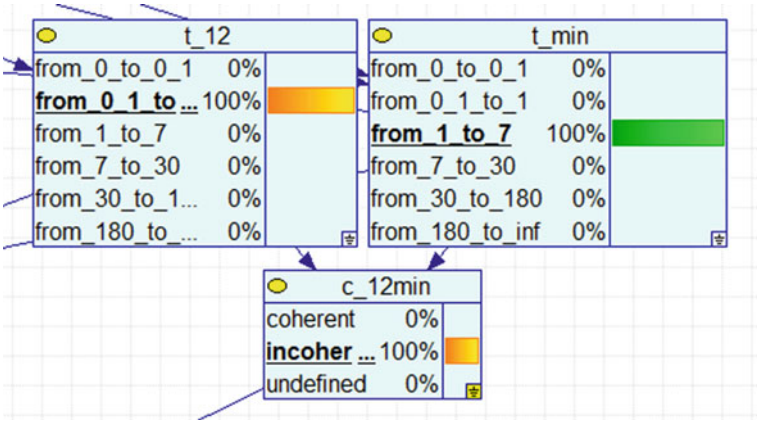

Fig. 5 Example of incoherent data (GeNIe)

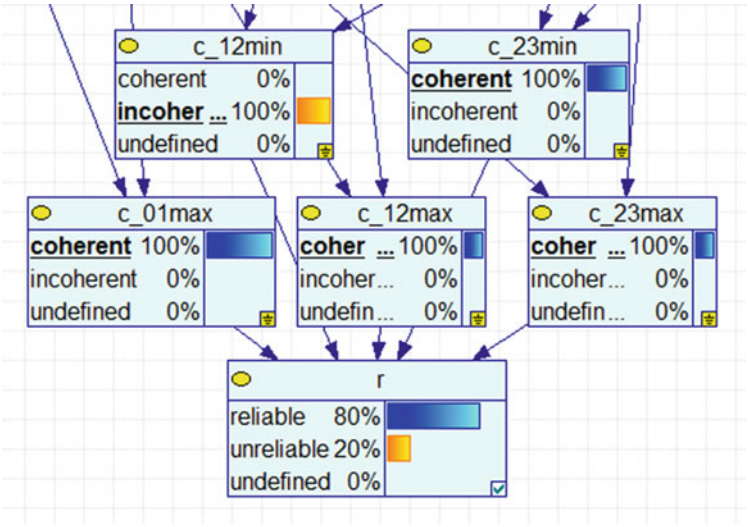

Fig. 6 Reliability estimation for the second respondent (GeNIe)

If the second respondent's data should be considered or excluded from the sample depends on the concrete research problem posed. If we want to use only the data without any contradictions or any uncertainties (all the data is coherent), then we take into account only the data from respondents 1, 4, 5, 7 and 10. Figure 7 shows the reliability estimation with the maximal degree of reliability.

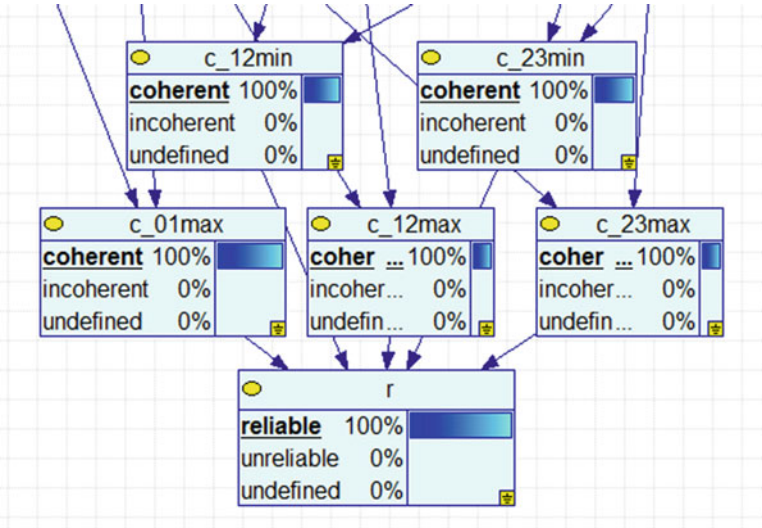

Fig. 7 Reliability estimation of respondent's coherent data (GeNIe)

6 Conclusions

We proposed a method of data coherence diagnosis of the risky behavior model with the data obtained from respondents. For more convenient use of the method software was developed and described. Example of the use of the method was also provided.

The more general cases of the coherence rate distribution (not only coherent-incoherent-undefined), different partition or unequal intervals can be considered.

This coherence diagnosis can be useful to eliminate not reliable respondents' data from the sample. Respondent reliability rate can be used as an analogue of lie scale in psychological test.

Acknowledgments This work was partially supported by the by RFBR according to the research project No. 16-31-00373.

References

1. Kabiri, P., Ghorbani, A.A.: Research on intrusion detection and response: a survey. Int. J. Netw. Secur. **1**(2), 84–102 (2005)
2. Mansour, M.M., Wahab, M.A.A., Soliman, W.M.: Bayesian networks for fault diagnosis of a large power station and its transmission lines. Electr. Power Compon. Syst. **40**(8), 845–863 (2012)

3. Xiong, G., Shi, D., Zhu, L., Duan, X.: A new approach to fault diagnosis of power systems using fuzzy reasoning spiking neural P systems. Math. Prob. Eng. **2013**, Article ID 815352, 13 pp. (2013). doi:10.1155/2013/815352
4. Zhu, Y., Huo, L., Lu, J.: Bayesian networks-based approach for power systems fault diagnosis. IEEE Trans. Power Deliv. **21**(2), 634–639 (2006)
5. Beshelev, S.D., Gurvich, F.G.: Mathematical and Statistical Methods of Expert Estimates, p. 263. Statistics, Moscow (1980)
6. Alonso, S., Chiclana, F., Herrera, F., Herrera-Viedma, E., Alcalá-Fdez, J., Porcel, C.: A consistency-based procedure to estimate missing pairwise preference values. Int. J. Intell. Syst. **23**, 155–175 (2008). doi:10.1002/int.20262
7. Tsyganok, V.V., Kadenko, S.V.: On sufficiency of the consistency level of group ordinal estimates. J. Autom. Inf. Sci. **42**(8), 42–47 (2010)
8. Muller, A., Mitchell, J., Crosby, R., Cao, L., Johnson, J., Claes, L., Zwaan, M.: Mood states preceding and following compulsive buying episodes: an ecological momentary assessment study. Psychiatry Res. **200**, 575–580 (2012)
9. Helouet, L., Marchand, H., Genest, B., Gazagnaire, T.: Diagnosis from scenarios. Discrete Event Dyn. Syst. Theory Appl. **24**(4), 353–415 (2013)
10. Suvorova, A.V., Tulupyev, A.L., Sirotkin, A.V.: Bayesian belief networks in problems of estimating the intensity of risk behavior. J. Russ. Assoc. Fuzzy Syst. Soft Comput. **9**(2), 115–129 (2014)
11. Suvorova, A.V.: Socially significant behavior modeling on the base of super-short incomplete set of observations. Inf.-Meas. Control Syst. **9**(11), 34–38 (2013)
12. Suvorova, A.V., Tulupyeva, T.V., Tulupyev, A.L., Sirotkin, A.V., Pashchenko, A.E.: Probabilistic graphical models socially significant behavior of the individual, taking into account incomplete information. Proc. SPIIRAS **3**(22), 101–112 (2012)
13. Paschenko, A.E., Tulupyev, A.L., Nikolenko, S.I.: HIV infection modeling based on the last episodes of risky behavior. Izvestiya Vuzov – Priborostroenie **11,** 33–34 (2006)
14. Fenton, K.A., Johnson, A.M., McManus, S., Erens, B.: Measuring sexual behaviour: methodological challenges in survey research. Sex. Transm. Infect. **77**, 84–92 (2001)
15. Sunner, L.E., Walls, C., Blood, E.A., Shrier, L.A.: Feasibility and utility of momentary sampling of sex events in young couples. J. Sex Res. **50**(7), 688–696 (2013)
16. Tulupyev, A.L., Nikolenko, S.I., Sirotkin, A.V.: Bayesian Networks: Logical-Probabilistic Approach, 607 pp. Nauka, St. Petersburg (2006)
17. Tulupyev, A.L., Sirotkin, A.V., Nikolenko, S.I.: Bayesian Belief Networks: Logical-probabilistic Inference in the Acyclic Directed Graph, 400 pp. University Press, St. Petersburg. (2009)
18. GeNIe&SMILE: Decisions Systems Laboratory, School of Information Sciences. University of Pittsburg. http://genie.sis.pitt.edu/