ENW TOP Grant 2018 Module 1

1 BASIC DETAILS

1a Title of the proposal

The Structural Complexity of Probabilistic Independence

1b Project Acronym

–

1c Application Module

Module 1

1d Details of the main applicant

Title: prof.dr.ir.
First name: Linda
Initials: L.C.
Last name: Gaag
Prefix: van der
Male/female: female
Address for correspondence:
Utrecht University
Department of Information and Computing Sciences
P.O. Box 80.089
3508TB Utrecht
The Netherlands
Preference for correspondence in English: no
Telephone: +31 (30) 253 4113
Cell phone: +31 (6) 230 59 221 (private)
Email: L.C.vanderGaag@uu.nl
Website(optional): –

1e Co-applicant

–

1f Group members involved in the proposed research

Members of the *Decision-support Systems* research group (*Algorithms* division), at Utrecht University:

Title: prof.dr.ir.
First name: Linda
Initials: L.C.
Last name: Gaag
Prefix: van der
Male/female: female

Position: full professor (1.0 fte)
Type of involvement: project coordinator, collaborator, daily supervisor, promotor
(*primary expertise*: Bayesian networks and their application, probabilistic inference, independence)

Title: dr.habil.
First name: Cassio
Initials: C.P.
Last name: Campos
Prefix: de
Male/female: male
Position: associate professor (UHD, 1.0 fte)
Type of involvement: contextual researcher
(*primary expertise*: probabilistic graphical models, machine learning)

Title: dr.ir.
First name: Janneke
Initials: J.H.
Last name: Bolt
Prefix: –
Male/female: female
Position: postdoc researcher (0.5 fte, started January 2018 for a period of four years)
Type of involvement: collaborator for the PhD students to be appointed
(*primary expertise*: Bayesian networks, independence)

Member of the *Algorithms & Complexity* research group (*Algorithms* division), at Utrecht University:

Title: prof.dr.
First name: Hans
Initials: H.L.
Last name: Bodlaender
Prefix: –
Male/female: male
Position: full professor (1.0 fte)
Type of involvement: collaborator for the postdoc researcher to be appointed
(*primary expertise*: algorithms design and analysis, computational complexity)

1g Scientific Summary

Probabilistic models are omnipresent, across a range of societal fields. The key to scalability of these models is *probabilistic independence.* In fact, the practicability of models like Markov random fields and Bayesian networks originates from their use of graph representations of independence. With real-world problems ever increasing in complexity and size, the current generation of models is reaching its limits, and advances in practicability are now called for. Such advances will most likely come from insights in independence. Despite existing studies, however, the structural complexity of independence is still largely unknown. In this setting, the proposed project will study the structural and decomposability properties of independence, to ultimately drive the design of a new generation of models with stronger representations of independence providing for more efficient inference.

Since independence relations typically are exponentially large in size, they are commonly represented by a small basis of statements; common properties of independence are then used for deriving any remaining independence statements. The project will investigate a range of set-theoretic operators on the bases of independence relations and thereby explore properties of structure and decomposability to foster representation by loosely coupled components. The project will further establish the computational complexities of various problems on independence, both in worst-case and amortised settings. These themes will be addressed, not just from the common perspective of a starting set of independence statements, but also from the novel perspective of sets of both independences and dependences, to more closely match a real-world application setting.

1h Abstract for layman

In veel gebieden in onze maatschappij worden probabilistische modellen gebruikt. Bij het opstellen van weersverwachtingen worden zulke modellen bijvoorbeeld gebruikt om de kans op extreme weersomstandigheden uit te rekenen, op grond waarvan dan eventueel een weersalarm wordt afgegeven. Ook in de geneeskunde worden probabilistische modellen gebruikt, bijvoorbeeld voor het bepalen van de kans op complicaties bij een specifieke behandeling.

Hoe complexer de toepassing, hoe meer tijd het doorrekenen van een probabilistisch model kost. Het gebruik van kennis van de onafhankelijkheden tussen de verschillende toevalsvariabelen in zo'n model zorgt er dan voor dat de rekentijd ingeperkt wordt. Echter, de toepassingen in onze maatschappij worden steeds groter en complexer, en de grenzen van de bruikbaarheid van de huidige modellen zijn in zicht. Het is daarom belangrijk dat nu al de grondslagen worden gelegd voor een volgende generatie van modellen die efficiënter kunnen worden doorgerekend. De sleutel tot bruikbaarheid voor steeds complexere toepassingen is hoogstwaarschijnlijk het begrip *onafhankelijkheid*: immers, hoe meer onafhankelijkheden kunnen worden weergegeven en benut, hoe efficiënter een model kan worden doorgerekend.

In dit project wordt het wiskundige begrip onafhankelijkheid vanuit verschillende (veelal informatica-) perspectieven bestudeerd. Enerzijds wordt de structuur van de onafhankelijkheidsrelatie in een kansverdeling onderzocht, met het oog op het opsplitsen ervan in redelijk losse delen die apart kunnen worden opgeslagen en gebruikt; het doel is om de opslag en het gebruik van die onafhankelijkheden effectiever te maken voor het doorrekenen van een probabilistisch model. Anderzijds wordt bestudeerd hoe kennis van *afhankelijkheden* tussen variabelen kan worden gebruikt om de onafhankelijkheidsrelatie scherper in kaart te brengen. Tot nu toe wordt in het onderzoek naar onafhankelijkheid eigenlijk alleen maar naar de onafhankelijkheden zelf gekeken, terwijl de afhankelijkheden ook een heleboel informatie geven: zo wordt het mogelijk om extra onafhankelijkheden te identificeren die niet worden tegengesproken door de gegeven afhankelijkheden en die misschien na verificatie ook kunnen worden benut. En tenslotte wordt in het project het complexiteitslandschap van het weergeven van en rekenen met onafhankelijkheden in kaart gebracht; hoe doel daarbij is om inzicht te krijgen in welke problemen wel en welke waarschijnlijk nooit echt efficiënt oplosbaar zullen zijn.

Hoewel op het gebied van probabilistische onafhankelijkheid al jaren onderzoek wordt gedaan, is de typische informatica-vraagstelling van efficiëntie nog weinig bestudeerd. Mede omdat het onderzoek behoorlijk fundamenteel van aard is en een degelijke achtergrond in zowel de wiskunde als de theoretische informatica vereist, is het aantal onderzoekers op het gebied redelijk beperkt. Verscheidene van de Europese onderzoekers zullen aan het voorgestelde project gelieerd worden door middel van concrete samenwerking, zodat onderzoeksresulaten snel internationaal zullen worden opgepakt. Het project zal uiteindelijk de weg banen naar een volgende generatie van probabilistische modellen waarmee complexe problemen sneller kunnen worden doorgerekend.

1i Physical Sciences research discipline

Research area	**Proposal applies to:**
Astronomy	
Computer Science	×
Mathematics	

1j Main field of research

Primary field of research:

16.30.00 Computer Science: Theoretical computer science

Secondary fields of research:

16.20.00 Computer Science: Software, algorithms, control systems
11.60.00 Mathematics: Probability theory, statistics

1k Keywords

Algorithms design and analysis; Tractability; Probabilistic modelling; Probabilistic independence.

1l Relevance to the 'Top sectors'

The project is not focused on a single societal sector, but has the potential to impact any sector in which (discrete) probabilistic models are applied. Given the applicant's long-standing experience with, and motivation from, the animal-production field, the Top sector *Agro & Food* is likely to be the first to benefit from advances driven by the project's results.

2 Research proposal

2a Overall aim and key objectives

Overall aim

Probabilistic models are omnipresent, and being developed for a yet wider range of real-world settings involving inherent uncertainty, among which are the medical, agricultural and ecological fields in which the applicant is active (see for example [6, 12, 15, 27, 29]). The key to the scalability of these models is *probabilistic independence.* Modern probabilistic models, such as Bayesian networks and Markov random fields, in fact, exploit graph representations of independence to arrive at feasible inference [7, 16, 17, 24], rather than build on general, (over-)simplifying assumptions. Despite existing studies of independence, however, its structural complexity is largely unknown. With real-world settings ever increasing in complexity and size, the current generation of models is reaching its computational limits, and further advances in practicability are now called for. Such advances are most likely to come from further insights in probabilistic independence, since stronger representations of independence will allow more efficient inference. In line with these observations, the proposed project will study properties of structure and decomposability of independence from various perspectives, to drive the design of a new, more powerful generation of models.

Scientific background

Probabilistic independence has been subject to multiple studies, from both a mathematics and a computer-science perspective (see for example [8, 24, 33]). An *independence relation* over a set of random variables V is a set of *triplets* $\langle A, B \,|\, C\rangle$ where $A, B, C \subseteq V$ are pairwise disjoint subsets of V with $A, B \neq \varnothing$; the set of all possible triplets is indicated by $V^{(3)}$. A triplet $\langle A, B \,|\, C\rangle$ essentially states that the sets of variables A and B are independent given the conditioning set C; relative to a (discrete) joint probability distribution Pr over V, the triplet thus states that $\Pr(A, B \,|\, C) = \Pr(A \,|\, C) \cdot \Pr(B \,|\, C)$ for all possible value combinations of A, B, C.

Well-known properties of probabilistic independence have been formulated as axiomatic systems to allow a study of independence without the numerical context involved (see for example [9, 13, 14, 22, 24, 28, 31, 34]). The most often studied system includes four axioms, called the *semi-graphoid axioms*. Any (ternary) relation $I \subseteq V^{(3)}$ closed under these axioms, is then called a *semi-graphoid independence relation* [13, 24]; in the sequel, the phrase independence relation is used to indicate a semi-graphoid independence relation, unless explicitly stated otherwise. The semi-graphoid system is known to be sound relative to the class of discrete probability distributions [9, 24], yet is not complete. In fact, it has been shown that probabilistic independence does not allow a finite axiomatisation [30]; a partial completeness result is known, however, which states that, for any two triplets, a probability distribution can be constructed with an independence relation that is composed of just these two triplets and the triplets derived from them by the semi-graphoid axioms [21]. Where the semi-graphoid system applies to the class of probabilistic independence relations in general, for various subclasses tailored systems have been formulated; for the independence relations of strictly positive probability distributions, for example, the semi-graphoid system has been extended with a fifth axiom to constitute the graphoid system [24].

Various computational problems on independence relations are being studied (see for example [1, 5, 11, 22, 35]). Within the proposed project, the focus will primarily be on the problem of convenient representation, as the representation used will ultimately influence the feasibility of probabilistic inference. Independence relations in general are typically exponentially large in the number of random variables involved [31, 32]. Representing such a relation by mere enumeration of its triplets therefore is not feasible in practice. By taking the four semi-graphoid axioms as rules for deriving new triplets, a more concise representation is arrived at by explicitly listing a small set of triplets, called a *basis*, and letting all other triplets be defined implicitly through these rules [32]; the basic idea is illustrated in Fig. 1.

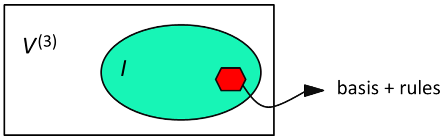


Figure 1: Representing an independence relation by a basis.

For convenient representation of semi-graphoid independence relations in general, two special types of basis have been proposed: an *elementary-triplet* basis is composed of triplets $\langle A, B \mid C\rangle$ with A, B singleton sets [20, 26], and a *dominant-triplet* basis is composed of triplets such that any remaining triplet can be derived directly from *one* triplet from this set [32]. The latter type of basis has received considerably more attention from the research community, since it is commonly thought to be smaller in size than an elementary-triplet basis and moreover provides for an efficient solution of the implication problem to be reviewed presently. Research so far has resulted in successive algorithms for computing a dominant-triplet basis from a given starting set of triplets, without the need to generate the full relation defined by this starting set [1, 2, 19, 32].

Entwined with the *representation problem* above, is the *implication problem* of establishing whether a given query triplet is implied by a given starting set of triplets. A commonly used approach to solving this problem is to establish, for a given starting set, a convenient basis representation, from which implication of the query triplet can be established efficiently. A dominant-triplet basis for example, provides for solving the implication problem for a given query triplet in linear time. As establishing a dominant-triplet basis is computationally challenging, this approach is beneficial especially if multiple triplets are queried from the same independence relation. Another approach to solving the implication problem is based on an imset representation of the starting set and the query triplet, which essentially results in a system of linear equations [22, 33]. If the system of equations has a solution, the query triplet is implied by the starting set. As the proposed project is centered on the more commonly employed basis representations, the imset approach will not be further discussed.

The semi-graphoid axioms of independence, and the idea of representing an independence relation by a basis, constitute the foundation on which the graph representations of independence used in modern probabilistic models are built (see for example [18, 25]). Current graph representations however cannot fully describe any semi-graphoid independence relation. In most real-world applications, in fact, various known independences escape explicit representation and therefore are not exploited upon probabilistic inference. While the theme of representing independence relations in graphical structures is not explicitly covered in the proposed project, the want of representational power of current representations is one of the project's motivations.

Research themes

The proposed study of probabilistic independence is focused on three research themes, for two PhD students and a postdoc researcher to be employed. The themes have been designed to collaboratively result in deep insight in the structural complexity of independence. Although the three themes allow independent studies, the more collaborative the research efforts are, the stronger the results are expected to be.

Theme 1: Decomposing independence relations (PhD)

> *Theme 1 pursues a representation of independence in which a relation is decomposed into weakly coupled components. The ultimate goal is to represent such components by separate bases and thereby arrive at a more concise overall representation. The theme will yield new insights in the structural properties of independence.*

A preliminary investigation of the decomposability of independence relations was conducted by the applicant [11, 19, 34]. An independence relation may embed other relations which have more structural regularity than the overall relation itself. For example, it may embed an *ascending*, or stable, independence relation such that for any included triplet $\langle A, B \,|\, C\rangle$ also all triplets $\langle A, B \,|\, C'\rangle$ with $C \subseteq C' \subseteq V \setminus (A \cup B)$ are included [20, 34]. Similarly, a *descending* relation may be embedded, that is, an independence relation such that for any included triplet $\langle A, B \,|\, C\rangle$ also all triplets $\langle A, B \,|\, C'\rangle$ with $C' \subseteq C$ are included [20]. These special independence relations have been studied by themselves, resulting in the formulation of additional properties of regularity. By taking these properties again as derivation rules, these relations can often be represented more compactly than semi-graphoid independence relations in general. For ascending relations, for example, two additional axioms have been formulated, allowing representation by a smaller basis than through the semi-graphoid system of axioms [11, 23, 34]. For exploiting the smaller bases of its embedded independence relations, and to thereby arrive at a more concise representation of a relation at large however, the overlaps among the embedded relations need be further investigated. For example, upon representing an embedded ascending relation by a separate basis, the remaining set of yet unrepresented triplets may not constitute an independence relation by itself. Extending this set to an independence relation by including some of the triplets from the stable relation will cause an overlap between the two parts of the original relation, which may result in a larger representation of the overall relation under study [11].

Independence relations can be decomposed also in other ways than through their embedded special relations. A relation can be partitioned, for example, into blocks of triplets with limited interaction through the derivation rules; an initial investigation of the feasibility of this idea has resulted in criteria under which a given triplet can never combine with any other triplet to yield new ones [19]. A relation can also be decomposed into subrelations defined by the sets of triplets with fixed conditioning sets; such a decomposition would for example support probabilistic reasoning in real-world settings in which evidence becomes available incrementally. The theme will explore various such decomposition approaches, along with the ensueing basis representations and their computation.

The starting point of study for the theme will be further investigation of a range of set-theoretic operators, among which are the intersection, union and set difference operators, defined on independence relations [3], as such operators will be instrumental for decomposing a relation into components for separate representation.

Theme 2: Enhancing independence representations with dependences (PhD)

Theme 2 pursues an enhanced representation of independence, in which a set of independences is combined with information of dependence. The ultimate goal is to arrive at better delineated representations of independence. The theme will yield new insights in inconsistency and lack of information of starting specifications of independence.

Thus far, studies into the representation of independence have been conducted from the perspective of a starting set of independence statements. Such starting sets then are taken to be perfect specifications of a semi-graphoid relation at hand, in the sense of being complete. In real-world settings however, it is more realistic to assume that the starting information is not perfect, since extracting independences from data through statistical tests or from experts by elicitation tends to be non-trivial and error-prone. The starting information may moreover be naturally composed of not just independence statements but of dependence information as well: dependences may arise from studying available data sets, and may in fact come to the fore in discussions with domain experts [10]. The proposed project will therefore explore the novel perspective of a starting set of both independence statements and dependence information.

The idea of enhancing representations of independence with information of dependence is to allow the identification of possible inconsistencies and/or lack of information of the specified independence relation from the overall information. An inconsistency arises if a particular ternary statement of non-overlapping sets of variables is implied to be a triplet of independence by the basis for the independence relation at hand, yet is also found to be a dependence by the available dependence information; lack of information occurs if such a statement is not implied by the basis for the independence relation and also is not found to describe a dependence. In an applications setting, problems of inconsistency and lack of information may be solved by further gathering efforts or by making assumptions, to thereby arrive at a better informed specification of the independence relation at hand.

To arrive at algorithms for identifiying properties of inconsistency and lack of information from starting sets, concise representations of probabilistic dependence need be investigated. Similar to the semi-graphoid system for independence, an axiomatic system has been developed for dependence [4]. Despite the long standing availability of this system, to the best of the applicant's knowledge, it has not been exploited as yet to foster the representation of a dependence relation by a small basis of statements. A starting point of study for the theme will therefore be further investigation of the axiomatic system for dependence and the representation of a dependence relation by a convenient type of basis.

Theme 3: Establishing the landscape of complexities (postdoc)

Theme 3 pursues insight in the complexities of a range of problems on independence relations. The ultimate goal is to identify the feasibility of different approaches to representing and reasoning about independence in probabilistic models.

Although it is common knowledge that independence relations can be exponentially large in their number of random variables and that representations by means of dominant triplets are more concise yet are costly to compute, surprisingly little is known about the exact representational and computational complexities involved. The only definite result at present is coNP-completeness of the implication problem for ascending independence relations [22]. A possible line of research is to establish the complexity classes for the problems of deciding whether a given arbitrary set of triplets is a minimal (with respect to set inclusion) basis for an independence relation defined by a starting set, and whether such a set is a minimum-sized basis. Another line of research is to establish the amortised complexity of establishing a dominant-triplet basis and solving the implication problem for a sequence of triplets. Representational issues pertain to, for example, the sizes of different types of basis with respect to the overall size of an independence relation at hand. Since the landscape of complexities is quite unexplored, the exact complexity issues to be addressed in the project will be decided upon jointly by the postdoc researcher to be appointed and the project team.

Research methodology
The three themes described above were designed primarily for foundational research, and are expected to result in theorems, proofs, data structures and algorithms. Each of the themes however, also provides for more experimental investigation, through implementation of algorithms and comparison of results on a range of starting sets. Within the broader scope of the project, some experimental investigation is foreseen, building upon an existing implementation [2], to lead to the construction of a website with state-of-the-art software and a collection of benchmark independence relations. The extent to which experimentation will be included in the studies of the separate themes of the proposed project, depends on the personal interests and capacities of the two PhD students and the postdoc researcher to be appointed.

Work plan

Researcher	**Year 1**	**Year 2**	**Year 3**	**Year 4**
PhD1	▨	▨	▨	▨
PhD2	▨	▨	▨	▨
Postdoc		▨	▨	▨

Schematic outline, over time, of the planned research:

PhD1:

Year 1: Getting acquainted with the field. Studying and extending related work on set-theoretic operators for independence relations. Designing a set-difference operator.

Year 2: Studying and characterising the overlap between an embedded ascending relation and the remainder of an independence relation. Identifying other embedded relations and studying their basis representations and possible overlaps.

Year 3: Identifying another promising approach to decomposition and detailing the ensueing representation.

Year 4: Rounding up and composing a thesis manuscript.

PhD2:

Year 1: Getting acquainted with the field. Studying related work on axiomatising dependence relations. Designing a basis representation for dependence.

Year 2: Defining inconsistency of a starting set of independence and dependence statements, and desiging algorithms for their detection and for their removal while adhering to the axioms of independence and dependence, respectively.

Year 3: Defining lack of information of a starting set of independence and dependence statements, and desiging algorithms for its detection and for extending the represented independence relation, without introducing inconsistencies and adhering to the axioms of independence.

Year 4: Rounding up and composing a thesis manuscript.

Postdoc:

Year 2: Getting acquainted with the field. Detailing the representational complexity of independency relations and their different types of basis. Establishing the complexity classes for decision problems pertaining to minimality properties of (arbitrary) bases.

Year 3: Studying the runtime complexities of existing algorithms for computing different types of basis and establishing the amortised complexities of subsequently solving the implication problem for a series of query triplets.

Year 4: Establishing the complexity classes of the implication problem for different types of independence relation.

The schematic research outline given above may be adjusted to the personal interests and capacities of the students and postdoc researcher employed, to include a stronger emphasis on experimentation. Furthermore, as with any foundational-research project in mathematics and computer science, the outlined plan may require adjustment as further insights with respect to feasibility are attained in the course of the project.

2b National and international cooperation

At her home institute, the applicant has a long-standing collaboration with prof.dr. H.L. Bodlaender (chair of the *Algorithms & Complexity* research group), with a range of joint publications. Professor Bodlaender has confirmed interest in the current project and will contribute his knowledge of complexity theory through collaboration with the postdoc researcher to be employed.

The applicant further has a strong network of contacts in her field of research at large, throughout Europe and, to a lesser extent, in the United States and Australia. The research community focusing on independence relations and their representation is rather small, possibly because the field is quite foundational and moreover requires proficiency in both mathematics and theoretical computer science. Relevant for the proposed project are the applicant's contacts with researchers at the University of Perugia (Italy), at Sapienza University of Rome (Italy), and at the Academy of Sciences in Prague (Czech Republic); the applicant recently spend a four-month sabbatical leave at the University of Perugia. Dr. M. Studený (Prague), dr. B. Vantaggi (Rome), and dr. M. Baioletti (Perugia) have confirmed interest in close collaboration within the project and to accommodate prolonged stays of the researchers involved in their respective home institutes.

2c Knowledge utilisation

Since use of the project's results will not be immediate for a specific field of application, but instead is expected to have a broad-scoped impact in five to ten years, knowledge utilisation during the project is sought through dissemination to the scientific community and to a more general public.

During the project, two workshops will be organised. The first of these will be held in the second year of the project, at the applicant's home institute, and is aimed at interaction of the project's researchers with a limited group of international colleagues addressing probabilistic independence; this workshop will foster exchange of ideas and initial collaborations in an informal setting. The second workshop will be held in the final year of the project, preferrably as a workshop at the Lorentz Center in Leyden. This workshop will be aimed at setting the research agenda for the next step in the design of a new generation of probabilistic models based on the results from the current project. Invited will be group of international researchers, not just from the field of probabilistic independence but also from graph theory. As a service to the research community, moreover, a website with state-of-the-art software and a collection of benchmark independence relations will be set up and maintained; this website will also contain leaflets, sets of slides, animations and other materials for educational purposes.

During the project, moreover, dissemination will be sought in popular scientific journals such as *Wetenschap in Beeld* and on websites such as *Scientias*, to reach out to the general public. Publications will showcase the power of probabilistic independence for efficiently solving non-trivial problems involving uncertainty. Following the applicant's positive experiences in the *Summer-break Symposium for Mathematics Educators* in 2016, similar stages for reaching mathematicians and computer scientists at all levels will be actively sought and pursued.

2d Literature references

References

[1] M. Baioletti, G. Busanello, B. Vantaggi (2009). Conditional independence structure and its closure: Inferential rules and algorithms. *International Journal of Approximate Reasoning*, 50: 1097–1114.

[2] M. Baioletti, G. Busanello, B. Vantaggi (2009). Closure of independencies under graphoid properties: some experimental results. In: *International Symposium on Imprecise Probability: Theories and Applications*, Durham, pp. 11 – 19.

[3] M. Baioletti, D. Petturiti, B. Vantaggi (2013). Qualitative combination of independence models. In: **L.C. van der Gaag** (editor). *Proceedings of the 14th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Lecture Notes in Artificial Intelligence 7958, Springer International Publishing, Heidelberg, pp. 37 – 48.

[4] R.R. Bouckaert (1994). Conditional dependence in probabilistic networks. In: P. Cheeseman, R.W. Oldford (editors). *Selecting Models from Data, Artificial Intelligence and Statistics IV*, Lecture Notes in Statistics 89, Springer-Verlag, Heidelberg, pp. 101 – 111.

[5] R.R. Bouckaert, R. Hemmecke, S. Lindner, M. Studený (2010). Efficient algorithms for conditional independence inference. *Journal of Machine Learning Research*, 11: 3453 – 3479.

[6] Th. Charitos, **L.C. van der Gaag**, S. Visscher, C.A.M. Schurink, P.J.F. Lucas (2009). A dynamic Bayesian network for diagnosing ventilator-associated pneumonia in ICU patients. *Expert Systems with Applications*, 36: 1249 – 1258.

[7] R.G. Cowell, A.P. Dawid, S.L. Lauritzen, D.J. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems.* Springer, Berlin.

[8] A.P. Dawid (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society B*, 41: 1 - 31.

[9] A.P. Dawid (2001). Separoids: A mathematical framework for conditional independence and irrelevance. *Annals of Mathematics and Artificial Intelligence*, 32: 335 – 372.

[10] **L.C. van der Gaag**, E.M. Helsper (2002). Experiences with modelling issues in building probabilistic networks. In: A. Gomez-Perez, V.R. Benjamins (editors). *Knowledge Engineering and Knowledge Management*, Springer, Berlin, pp. 21 – 26.

[11] **L.C. van der Gaag**, S. Lopatatzidis (2017). Exploiting stability for compact representation of independency models. In: A. Antonucci, L. Cholvy, O. Papini (editors). *Proceedings of the 14th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Lecture Notes in Artificial Intelligence 10369, Springer, Berlin, pp. 104 – 114.

[12] P.L. Geenen, **L.C. van der Gaag**, W.L.A. Loeffen, A.R.W. Elbers (2011). Constructing naive Bayesian classifiers for veterinary medicine: a case study in the clinical diagnosis of classical swine fever. *Research in Veterinary Science*, 99: 64 – 70.

[13] D. Geiger, A. Paz, J. Pearl (1991). Axioms and algorithms for inferences involving probabilistic independence. *Information and Computation*, 91: 128 – 141.

[14] D. Geiger, J. Pearl (1993). Logical and algorithmic properties of conditional independence and graphical models. *The Annals of Statistics*, 21: 2001 – 2021.

[15] E.M. Helsper, **L.C. van der Gaag** (2007). Ontologies for probabilistic networks: a case study in the oesophageal-cancer domain. *The Knowledge Engineering Review*, 22: 67 – 86.

[16] M.J. Jordan (2004). Graphical models. *Statistical Science*, 19: 140 – 155.

[17] D. Koller, N. Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques.* MIT Press.

[18] S.L. Lauritzen (1996). *Graphical Models*, Clarendon Press.

[19] S. Lopatatzidis, **L.C. van der Gaag** (2015). Concise representations and construction algorithms for semi-graphoid independency models. *International Journal of Approximate Reasoning*, 80: 377–392.

[20] F. Matúš (1992). Ascending and descending conditional independence relations. *Proceedings of the Eleventh Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pp. 189–200.

[21] F. Matúš (2004). Towards classification of semigraphoids. *Discrete Mathematics*, 277: 115 – 145.

[22] M. Niepert, B. Sayrafi, M. Gyssens, D. Van Gucht (2013). The conditional independence implication problem: A lattice-theoretic approach. *Artificial Intelligence*, 202: 29 – 51.

[23] M. Niepert, D. Van Gucht, M. Gyssens (2010). Logical and algorithmic properties of stable conditional independence. *International Journal of Approximate Reasoning*, 51: pp. 531–543.

[24] J. Pearl (1988). *Probabilistic Reasoning in Intelligent Systems. Networks of Plausible Inference.* Morgan Kaufmann, Palo Alto.

[25] J. Pearl, D. Geiger, T.S. Verma (1989). Conditional independence and its representations. *Kybernetika,*, 25: 33 – 44.

[26] J.M. Peña (2017). Representing independence models with elementary triplets. *International Journal of Approximate Reasoning*, 88: 587 – 601.

[27] R.F. Ropero, S. Renooij, **L.C. van der Gaag** (2018). Discretization methods for learning Bayesian network-based classifiers from environmental data. *Ecological Modelling*, 368: 391 – 403.

[28] W. Spohn (1980). Stochastic independence, causal independence and shieldability. *Journal of Philosophical Logic*, 9: 73 – 99.

[29] W. Steeneveld, **L.C. van der Gaag**, W. Ouweltjes, H. Mollenhorst, H. Hogeveen (2010). Discriminating between true-positive and false-positive clinical mastitis alerts from automatic milking systems. *Journal of Dairy Science*, 93: 2559 – 2568.

[30] M. Studený (1992). Conditional independence relations have no finite complete characterization. In: S. Kubík, J.Á. Vísek (editors). *Information Theory, Statistical Decision Functions and Random Processes.* Kluwer, Amsterdam, pp. 377–396.

[31] M. Studenỳ (1997). Semigraphoids and structures of probabilistic conditional independence. *Annals of Mathematics and Artificial Intelligence*, 21: 71–98.

[32] M. Studený (1998). Complexity of structural models. *Proceedings of the Joint Session of the 6th Prague Conference on Asymptotic Statistics and the 13th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pp. 521–528.

[33] M. Studený (2005). *Probabilistic Conditional Independence Structures.* Springer Verlag, London.

[34] P. de Waal, **L.C. van der Gaag** (2004). Stable independence and complexity of representation. In: M. Chickering, J. Halpern (editors). *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence.* AUAI Press, Arlington, pp. 112–119.

[35] S. Wong, C. Butz, D. Wu (2000). On the implication problem for probabilistic conditional independency. *IEEE Trans. Systems, Man, and Cybernetics, Part A: Systems and Humans*, 30: 78 – 805.

2e Data management

Although the three themes to be covered by the proposed project have been designed primarily for foundational research, some experimental investigation is foreseen. A website will be set up to make state-of-the-art-algorithms publicly available. Through the website, also benchmark independence relations will be made available by means of their starting sets. Each of the independence relations will be well documented, in terms of their embedded relations and in terms of the sizes of different types of (intermediate) basis. As the thus provided information is intended for re-use by others in the research community, the project team will design an appropriate set-up for the website as well as a detailed documentation protocol in close consultation with the project's international collaborators.

3 Cost estimates

3a Budget

a.	appointment of research personnel	
	PhD 4 years Postdoc 3 years	2 1
b.	additional travel budget	€17,000
c.	project-related equipment	€0
d.	other project-related activities	€8,000
	Total b, c, d	€25,000

In the course of the proposed project, the applicant hopes to form a collaborative European consortium in the research field, for follow-up studies. To this end, she will promote close collaboration for the Netherlands-based project team with the field's key researchers throughout Europe. Extra travel budget is requested therefore, for facilitating prolonged stays of the project's researchers in the Czech Republic and in Italy, and for the European key researchers to spend time with the group in Utrecht. The costs for the three European researchers affiliated with the proposed project, for a one-month stay in the Netherlands each, are estimated at €10,000. A budget of €7,000 is requested in addition to the standard bench fees for the two PhD students and the postdoc researcher, to allow prolonged stays in Italy and in the Czech Republic.

During the project, two workshops will be organised. The first of these workshops will be rather informal, with at most five participants from abroad and some five participants from Utrecht; the costs involved in the organisation of this three-day workshop are estimated at €4,000, involving the costs of travel and accommodation of the participants from abroad. To cover the costs of the second, five-day workshop, also a budget of €4,000 is requested. For the organisation of this workshop, the applicant will submit a proposal to the Lorentz Center in Leyden (see `http://www.lorentzcenter.nl/infoorg.php`). If the workshop will not be organised at the Lorentz Center, additional funding will be sought to cover all costs of the expected ten to fifteen particpants.

3b Number of words used

Section	**number of words used**	**maximum number of words allowed**
1g	249	250
1h	447	500
2a+2b+2c	3161	4000
2c	305	500
4f	241	250
4	1807	2000

3c Grant applications

None.

3d Open positions

None.

4 CURRICULUM VITAE (max: 2000 words)

4a Personal details

Title, inital, first name, surname: prof.dr.ir. L.C. (Linda) van der Gaag
Male/female: female
Date and place of birth: July 23rd, 1959, Delft, the Netherlands
Nationality: Dutch
PhD date: September 26th, 1990

4b Current Employment

Full professor, permanent position (1.0 fte), Utrecht University, the Netherlands.

4c Work experience since completing your PhD

Position	Period	fte	Type of position	Institute
Assistant professor	1989 – 1995	1.0	permanent	Utrecht University
Associate professor	1995 – 2000	1.0	permanent	Utrecht University
Full professor	2000 - now	1.0	permanent	Utrecht University
Honorary professor	2000 – 2005	–	temporary	University of Aberdeen, Scotland

Since her appointment as a full professor in 2000, the size of the applicant's research group has varied between 3 and 27 fte, with the smallest number of fte in 2017 after the applicant had served as head of department for more than three years (November 2013 – January 2017). As of December 1st, 2017, the research group is once more expanding.

4d Work experience in months spent since completing your PhD

Experience	Number of months
Research activities	110.35
Education	97.3
Care or sick leave	37
Management tasks	86.35
Other, please specify	-
Total	331

Calculation of person-years of research activities (in months):

i. October 1990 - June 1995, Assistant professor, 1 fte; 40% research, 50% education, 10% management:
57 months * 0.4 research = 22.8 months of research

ii. July 1995 – April 2000, Associate professor, 1 fte; 40% research, 40% education, 20% management:
58 months * 0.4 research = 23.2 months of research

iii. May 2000 – January 2008, Full professor, 1 fte; 40% research, 30% education, 30% management:
93 months * 0.4 research = 37.2 months of research

iv. February 2008 – September 2010, Sick leave

v. October 2010 – October 2013, Full professor, 1 fte; 35% research, 25% education, 40% management:
37 months * 0.35 research = 12.95 months of research

vi. November 2013 – December 2016, Full professor / Head of departement, 1 fte; 20% research, 15% education, 65% management:
38 months * 0.2 research = 7.6 months of research

vii. January 2017 – May 2017, Sick leave

viii. June 2017 – April 2018, Full professor, 1 fte; 60% research, 25% education, 15% management:
11 months * 0.6 research = 6.6 months of research

The reported percentages are (informed) approximations. Only sick leaves of more than two consecutive months have been reported in the table above.

4e Academic staff supervised

	Promotor	**Co-promotor**	**Supervisor**
PhDs (*ongoing*)	M.T. Rietbergen (defense expected in 2018) K.L. Sadowski (defense expected in 2018)		
PhDs (*successfully completed*)	J.H. Bolt (2008) P.A.N. Bosman (2003) Th. Charitos (2007) V.M.H. Coupé (2000) M.M. Drugan (2006) J. Kwisthout (2009) S. Renooij (2001) D. Sent (2005) W. Steeneveld (2010) S.P.D. Woudenberg (2015)	R.R. Bouckaert (1995)	
Subtotal	12	1	–
Postdocs			J.H. Bolt M.M. Drugan S.F. van Dijk P.L. Geenen E.D. de Jong M. Oud
Subtotal			6
Programmers			M. Frasca A.-J. de Groote A. van IJzendoorn M.M. Schrage
Subtotal			4

4f Brief summary of research

The applicant's research field at large concerns the design and analysis of probability-based decision-support systems that are aimed at helping human decision makers take optimal decisions for problems pervaded with uncertainty. In many fields of society, experts are called upon to solve problems which are of an increasing complexity on the one hand and the possible solutions of which may have far-reaching consequences on the other hand; tailored, high-quality support from computer-based systems then is called for. Building upon concepts and techniques from probability theory and statistics, the applicant's main research goal is to advance the practicability of such systems for real-world problem domains.

The applicant's research activities range from addressing foundational challenges to developing applications. Her applications field of choice is the biomedical field, as the omnipresence of uncertainty makes this field particularly amenable to probability-based decision support; she has realised a number of real-world applications in this field in close collaboration with domain experts. Many of the applicant's foundational research activities are driven by questions that arise in such practical settings. Her foundational research efforts have resulted, among other lines of research, in algorithms for studying the sensitivity of the output of Bayesian networks in terms of inaccuracies in their parameters. Studying various problems on Bayesian networks and thereby encountering computational limits, have made her realise that further advances in the practicability of probabilistic graphical models can only be achieved through stronger representations of independence.

4g Other academic activities

Conference organisation and other involvements
In addition to a range of workshops, the applicant has organised the three main conferences in her research field:

- *Uncertainty in Artificial Intelligence* (UAI'07), in Vancouver, Canada: Conference co-chair with R. Parr;
- *Symbolic and Quantitative Approaches to Reasoning with Uncertainty* (ECSQARU'13), in Utrecht, the Netherlands: Conference chair;
- *Probabilistic Graphical Models* (PGM'14), in Utrecht, the Netherlands: Conference co-chair with A. Feelders.

As a senior researcher in her field, and in the field of Artificial Intelligence at large, the applicant partakes in the programme committees of all main conferences. By her own choice, the applicant is no longer a member of the editorial board of a journal, yet serves as a guest editor for some journals in her field (such as *IEEE Transactions on Knowledge and Data Engineering*, *Artificial Intelligence in Medicine*, and *International Journal of Approximate Reasoning*).

Research output
The following table lists the research output of the applicant, per publication category. Publications in conferences and workshops are included only if peer reviewed, and only publications in English are reported.

Publication	**Number**
Books, proceedings	8
Conference and workshop papers	148
Journal papers	40
Total	196

Within the computer-science research field in general, there is a strong focus on conference publications. Conference publications essentially are full papers and are assigned more value than in most other fields of science; publications at top conferences in fact are often considered of higher value than journal publications. The applicant has the policy to add her name as an author on a paper only if she had a considerable contribution to its scientific contents. The applicant further holds as a general rule that her PhD students are mentioned as first authors.

Dissemination
As any senior researcher in computer science, the main applicant has given multiple talks at conferences, workshops and other meetings. She was invited to give plenary talks at the following events (since 2012):

- Alan-Turing Year Celebration, 2012; talk broadcasted on regional television; title of the talk: *Kennissystemen in de Geneeskunde: Van Mythe naar Gemeengoed ?* (in Dutch), Almere, the Netherlands;
- BVPA'12 (Meeting of the British Veterinary Poultry Association); title of the talk: *An Early Warning System for Low-pathogenic Avian Influenza*, Harrogate, UK;
- ISIPTA'13 (International Symposium on Imprecise Probability: Theories and Applications); title of the talk: *Recent Advances in Sensitivity Analysis of Bayesian Networks*, Compiègne, France;

- ORAFM'14 (Conference on OR in Agriculture and Forest Management); title of the talk: *Expert Probabilities in Bayesian Networks for Early Warning*, Lleida, Spain.

The applicant has delivered multiple international master classes at PhD level, on the foundations and application of Bayesian networks. In addition, she featured at the Summer-break Symposium for Educators of Mathematics in 2016 in the Netherlands, with the talk *Van de Predikant Bayes naar Bayesiaanse Netwerken* (in Dutch).

4h Scholarships, grants and prizes

Grants

As PI	**Amount in € for UU**	**Year of award**
Pionier (NWO)	454,000	2000
Promundi (NWO)	121,000	2007
Multi (NWO)	210,000	2010
BioBayes (NWO)	207,000	2012
Subtotal	992,000	
As co-Applicant		
Competent GAs (NWO)	134,500	2000
TimeBayes (NWO)	139,400 (544,400 for consortium)	2001
EPIZONE (EU,CIDC)	312,900	2005
AniBioThreat (EU, CVI)	222,400	2011
ESBL (1Health4Food)	355,000 (4,137,000 for consortium)	2013
Poultry sector (1Health4Food)	135,000 (335,000 for consortium)	2014
Subtotal	1,299,200	
Total	2,291,200	

The table lists the larger grants attained by the applicant since 2000, that is, grants involving funding for at least one PhD student or postdoc researcher. Some 60% of the reported total amount of funding from these grants was received for applications-oriented research, mostly within the animal-production and human-health sectors. Smaller amounts of funding acquired from external parties are not included in the table above.

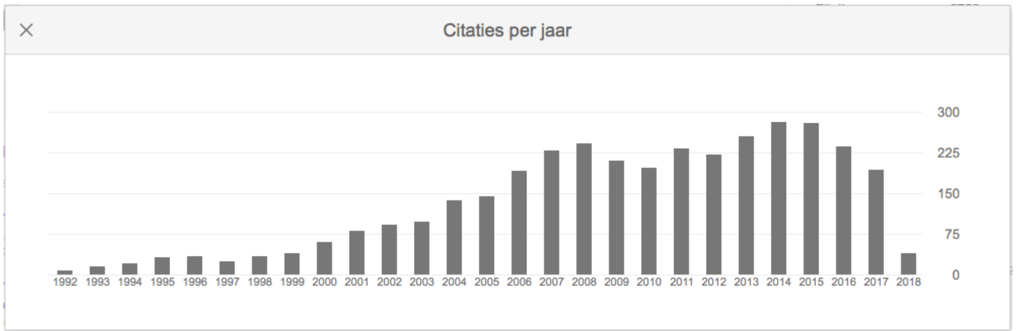
Prizes
Relevant to the current proposal is the *best-paper award* won by the applicant and her student at the ECSQARU'15 conference:

> S. Lopatatzidis, L.C. van der Gaag (2015). Computing concise representations of semi-graphoid independency models. In: S. Destercke, Th. Denoeux (editors). *Proceedings of the 13th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Lecture Notes in Artificial Intelligence 9161, Springer, Berlin, pp. 290 - 300.

In August 2007, the applicant was awarded an ECCAI (European Association for Artificial Intelligence) Fellowship for *Pioneering Work in the Field of Artificial Intelligence and Outstanding Service for the European Artificial Intelligence Community.*

4i Other relevant information

The applicant's h-index equals 32 (Google Scholar, April 27th, 2018), with an i10-index of 65 and with 3750 citations. The development of her citations over time is as follows:



The dip in citations after 2008 is explained by the applicant's absence from research for a period of almost three years due to medical reasons; the citation dip after 2015 is explained by her absence from research meetings due to her role as head of departement and personal circumstances.

The applicant has compared her h-index with those of the members of the programme committees of the biennial probabilistic graphical models (PGM) conference. This conference matches her research profile in a broad sense, and its programme committee is composed of representative members of the associated research community. Her h-index ranges among the top 10–15% of those of the committee members of previous events.

4j Five top-publications and five most relevant

The applicant considers the following papers as her five top-publications (listed in chronological order). These publications are not the most referenced ones, but rather publications that laid the foundations of new lines of research.

- M.J. Druzdzel, **L.C. van der Gaag** (1995). Elicitation of probabilities for belief networks: combining qualitative and quantitative information. In: Ph. Besnard, S. Hanks (editors). *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, pp. 141 – 148.
- V.M.H. Coupé, **L.C. van der Gaag** (2002). Properties of sensitivity analysis of Bayesian belief networks. *Annals of Mathematics and Artificial Intelligence*, 36: 323 – 356.
- **L.C. van der Gaag**, H.L. Bodlaender, A. Feelders (2004). Monotonicity in Bayesian networks. In: M. Chickering, J. Halpern (editors). *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*, AUAI, Arlington, pp. 569 – 576.
- A. Feelders, **L.C. van der Gaag** (2006). Learning Bayesian network parameters under order constraints. *International Journal of Approximate Reasoning*, 42: 37 – 53.
- J.H.P. Kwisthout, H.L. Bodlaender, **L.C. van der Gaag** (2010). The necessity of bounded treewidth for efficient inference in Bayesian networks. In: H. Coelho, R. Studer, M. Wooldridge (editors). *Proceedings of the 19th European Conference on Artificial Intelligence*, IOS Press, Amsterdam, pp. 237 – 242.

The five publications most relevant for the current research proposal are (in chronological order):

- **L.C. van der Gaag**, J-J.Ch. Meyer (1998). Informational independence: models and normal forms. *International Journal of Intelligent Systems*, 13: 83 – 109.

- P. de Waal, **L.C. van der Gaag** (2004). Stable independence and complexity of representation. In: M. Chickering, J. Halpern, editors. *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*, AUAI Press, Arlington, Virginia, pp. 112 – 119.

- P. de Waal, **L.C. van der Gaag** (2005). Stable independence in perfect maps. In: F. Bacchus, T. Jaakkola (editors). *Proceedings of the Twenty-first Conference on Uncertainty in Artificial Intelligence*, AUAI Press, Corvallis, pp. 161 – 168.

- S. Lopatatzidis, **L.C. van der Gaag** (2017). Concise representations and construction algorithms for semi-graphoid independency models. *International Journal of Approximate Reasoning*, 80: 377 – 392.

- **L.C. van der Gaag**, S. Lopatatzidis (2017). Exploiting stability for compact representation of independency models. In: A. Antonucci, L. Cholvy, O. Papini (editors). *Proceedings of the 14th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Lecture Notes in Artificial Intelligence 10369, Springer, Berlin, pp. 104 – 114.

By submitting this document I declare that I satisfy the nationally and internationally accepted standards for scientific conduct as stated in the Netherlands Code of Conduct for Scientific Practice 2012 (Association of Universities in the Netherlands).