Research Statement

Timothy van Bremen
Research Fellow, National University of Singapore
https://www.comp.nus.edu.sg/~tvanbr/

My research lies in the areas of **statistical-symbolic artificial intelligence** and **data management under uncertainty**. There is a growing need for hybrid systems that combine the successes of recent developments in neural and statistical AI, with classical symbolic techniques whose behaviour can be rigorously verified and explained. I study the theory underpinning the *intersection between statistical methods and symbolic AI*, and design and develop practical tools to tackle related problems. My research work has led to the release of at least **three** open-source tools, and straddles the boundary between theory and practice, as well as between different academic communities. I have accordingly divided my work into two key themes, on the basis of application area.

Theme 1: Inference in Statistical-Symbolic Models

In the past few decades, the world has witnessed significant growth in data collected in an automated fashion, through agents like machine learning systems or sensor networks. Common to many of these settings is **two** key attributes:

1. First, the collected data is often *uncertain*—for example, an NLP system extracting facts from text on the Internet may be wrong with some probability, either due to errors in the underlying model, or inaccuracies in the text itself.
2. Second, the data is often inherently *relational* in nature. For instance, an NLP system may extract the information that Alonzo Church is the doctoral advisor of Alan Turing, where "is the advisor of" here is the underlying relation (e.g., `is_advisor_of(church,turing)`).

It is therefore desirable to develop systems that can reason about and learn from such data in a scalable, transparent, and principled way. *Statistical-symbolic* models aim to provide a unified formalism for representing probabilistic, relational knowledge, by combining probabilistic graphical models like Markov random fields and Bayesian networks with relational representations, typically based on first-order logic or logic programming. Such models include Markov logic networks, Probabilistic Soft Logic, and probabilistic programming languages like ProbLog. These models all have the advantage of being **explainable by design**, since their semantics are in terms of human-readable logical formulas.

Lifted Inference Inference in many statistical-symbolic models like the ones listed above can be reduced to *first-order model counting*, a fundamental task in logic that can be viewed as a sort of "assembly language" for probabilistic inference. It generalizes the *model counting* problem—determining the number of satisfying assignments of a propositional formula—to the first-order logic setting, and is known to be #P-hard (the counting analogue of NP-hardness). A key advantage of casting inference problems in this way is that it allows for a clean separation between *domain*—for example, people in a social network—and *rules*—for example, the fact that people who are friends with smokers are more likely to smoke themselves. Inference algorithms that are able to exploit the symmetries arising from this separation, and scale in time polynomial (rather than exponential) in domain size are called *lifted* inference algorithms. Such inference algorithms can scale to domain sizes **many orders of magnitude** larger than classical non-lifted algorithms—indeed, our results typically show a **well over 1000× speed-up in inference time** [UAI-21; AAAI-22]. Thus, lifted inference represents a *fundamental shift* from classic inference methods for statistical-symbolic models, and such techniques are essential to be able to reason efficiently about relational data at Internet scale (i.e., billions of relations).

Practical Lifted Inference and Sampling Motivated by these applications, we have worked on developing a fast system for domain-lifted inference in statistical-symbolic models through first-order model counting, dubbed FASTWFOMC [UAI-21], that builds on dynamic programming and algorithmic techniques to outperform the state of the art in this area. In addition to its applications in probabilistic inference, it has also enjoyed success in applications focused on conjecturing and computing novel integer sequences in combinatorics, where we used it to find as-yet **undiscovered recurrences for sequences in the Online Encyclopedia of Integer Sequences** [ILP-21]. In a separate line of work which won **runner-up for best student paper** at KR 2021, we expanded the fragment of first-order logic (and thus, class of statistical-symbolic models) known to admit lifted inference algorithms [KR-21; AIJ-23]. Finally, we have also shed light on analogous problems in *sampling*—given a statistical-symbolic model, drawing samples from it in time polynomial in the domain size—as well as related questions [AAAI-22; IJCAI-21].

Propositional Inference On the other hand, for certain statistical-symbolic models it can be proved that no lifted inference algorithm can exist. In these cases, we are left with reducing the inference task to the #P-hard problem of *propositional* model counting (the counting analogue of the well-known Boolean satisfiability problem) on a "grounded" form of the input. We have worked on developing exact propositional model counting algorithms that can effectively exploit symmetries that tend to occur in practice in propositional formulas like these [AAAI-21]. In a complementary direction [IJCAI-20], we show how to leverage approximate propositional model counters for inference in a more efficient way than the classic ground-and-solve paradigm, by exploiting symmetries in the input *weights*, yielding a novel angle on this problem.

Theme 2: Data Management under Uncertainty

I am also interested in *data management under uncertainty*, and have focused on gaining an understanding of the theoretical principles behind *(tuple-independent) probabilistic databases*. Probabilistic databases are a simple but powerful formalism for incorporating uncertainty into classic relational databases. In this model, each tuple appearing in a relational database is annotated with an independent probability value, that intuitively describes the likelihood of that tuple being true. The primary inference task in this setting is known as *probabilistic query evaluation* (PQE): given a Boolean (true/false) query Q on a probabilistic database D, determine the probability that Q evaluates to true on a random database subinstance $D' \subseteq D$ obtained by including each tuple with its corresponding probability. Despite its straightforward semantics, the PQE problem is notoriously intractable ($\#\mathsf{P}$-hard) for all but the simplest of queries. In fact, the aforementioned first-order model counting problem is known to be a special case of PQE, thus highlighting the hardness of the problem.

Ontology-Mediated Querying I have worked on developing a practical system, ONTO2PROBLOG, for querying certain probabilistic databases in the presence of an *ontology*: deterministic knowledge specified in some logical language, expressing global rules—say, that every student has precisely one advisor, or every class has at most fifty students [CIKM-19; KI-20]. The system works by employing “combined rewritings” to reduce the queries to inference in a probabilistic logic program, allowing us to take advantage of the inference techniques for statistical-symbolic models discussed above.

Approximation Algorithms More recently, since starting my postdoc at the National University of Singapore, I have become interested in other facets of the PQE problem. Due to its inherent intractability, the vast majority of work on PQE has only considered identifying tractable cases of the problem measured in terms of the *database* size alone, ignoring possible exponential factors in the query length which can make the problem infeasible in practice. In a recent breakthrough [PODS-23], we showed that, in effect, any conjunctive query that can be efficiently evaluated on classical (deterministic) databases can also be *approximated* in randomized polynomial time on probabilistic databases, measured in *both* the query and database size (“combined complexity”), subject to only one condition: that no relation name appears more than once in the query (“self-join-freeness”). In particular, our approach gives PAC-style (ϵ, δ)-guarantees, ensuring that the output lies within a $(1 + \epsilon)$-multiplicative factor of the correct answer with probability at least $(1 - \delta)$. All existing approximation algorithms here either suffer from an exponential dependence on query size, rendering them of limited practical utility, or lack strong theoretical guarantees on their accuracy. This discovery has opened up the **opportunity to revisit** many of the landmark results in probabilistic databases, both from the angle of combined complexity and approximations.

Hardness and Network Reliability In newer work [ICDT-24], we expand our study on combined approximations by considering various natural restrictions on the structure of the database instance, and prove a variety of approximability as well as hardness of approximation results. In the process, we show that our earlier PODS 2023 result is *optimal*—that is, dropping the self-join-free restriction on the query means that provably no polynomial-time approximation algorithm can exist unless $\mathsf{RP} = \mathsf{NP}$. We also exhibit a surprising connection to the *two-terminal network reliability* problem, a fundamental problem studied in networks and operations research. We show the existence of a fully polynomial-time randomized approximation scheme (FPRAS) for this problem on directed acyclic graphs, **resolving a long-standing open question** in this area. The result has important implications for verifying reliability and resilience in domains such as power and transportation networks, which I look forward to exploring more deeply.

Future Vision

Since statistical-symbolic models and probabilistic databases are inherently interpretable, the need for such models will only grow. As a result, the research themes mentioned here have given me many exciting paths to explore in the years to come. I highlight just two of the several different research directions I will investigate below, which contribute to, and build on, ideas in both foundational research and practice.

Data Structures for Approximate Inference I will study the design and implementation of *data structures for approximate inference and sampling* with strong theoretical guarantees. Work in the field of *knowledge compilation* has studied classes of models (in AI) and queries (in databases) that can be “compiled” to compact data structures

(OBDDs, SDDs, d-DNNFs, etc.) that admit tractable *exact* inference and sampling. But what does the picture looks like for approximations? For example, recent work at STOC 2021 showed that the class of *structured DNNF* circuits allows for approximate counting of satisfying assignments in randomized polynomial time. Our PODS 2023 paper has made use of this to approximately evaluate queries in probabilistic databases, but we have only scratched the surface of its implications for AI, in terms of the possibility of defining a far **larger** class of statistical-symbolic models that admit scalable *approximate* inference. I would also like to better understand how this result connects to more sophisticated probabilistic queries (e.g., marginal MAP), as well as weight and structure learning. In addition, the design and development of comprehensive practical tooling here is a wide open gap that I am eager to bridge, so that such techniques can see widespread real-world use.

Theory and Tooling for Probabilistic Knowledge Graphs I am also interested in connecting the themes above with problems in open-world querying of probabilistic knowledge graphs, which is closely related to probabilistic databases, and has seen great success with applications in areas such as computational biology, computational social science, and recommender systems. On the one hand, much of the work in the machine learning community on probabilistic knowledge graphs has revolved around graph embeddings, but such methods typically lack interpretability, as well as comprehensive theoretical guarantees on their runtime and accuracy. On the other hand, the database theory community has developed a principled theory for querying uncertain data, but consideration of the open-world setting from a practical perspective is comparatively lacking. I therefore see an exciting opening here to bring together results from both communities to develop a **highly scalable system** for reasoning and learning in probabilistic knowledge graphs, with a rigorous theoretical underpinning. This direction also opens up the possibility of exploring a wealth of interesting connections to topics in neurosymbolic AI.

References

[AAAI-21] Timothy van Bremen, Vincent Derkinderen, Shubham Sharma, Subhajit Roy, and Kuldeep S. Meel. "Symmetric Component Caching for Model Counting on Combinatorial Instances". In: *AAAI Conference on Artificial Intelligence (**AAAI**)*. 2021, pp. 3922–3930.

[AAAI-22] Yuanhong Wang, Timothy van Bremen, Yuyi Wang, and Ondřej Kuželka. "Domain-Lifted Sampling for Universal Two-Variable Logic and Extensions". In: *AAAI Conference on Artificial Intelligence (**AAAI**)*. 2022, pp. 10070–10079.

[AIJ-23] Timothy van Bremen and Ondřej Kuželka. "Lifted Inference with Tree Axioms". In: *Artif. Intell. (**AIJ**)* 324 (2023), p. 103997.

[CIKM-19] Timothy van Bremen, Anton Dries, and Jean Christoph Jung. "Ontology-Mediated Queries over Probabilistic Data via Probabilistic Logic Programming". In: *ACM International Conference on Information and Knowledge Management (**CIKM**)*. 2019, pp. 2437–2440.

[ICDT-24] Antoine Amarilli, Timothy van Bremen, and Kuldeep S. Meel. "Conjunctive Queries on Probabilistic Graphs: The Limits of Approximability". In: *International Conference on Database Theory (**ICDT**)*. 2024. Forthcoming.

[IJCAI-20] Timothy van Bremen and Ondřej Kuželka. "Approximate Weighted First-Order Model Counting: Exploiting Fast Approximate Model Counters and Symmetry". In: *International Joint Conference on Artificial Intelligence (**IJCAI**)*. 2020, pp. 4252–4258.

[IJCAI-21] Yuanhong Wang, Timothy van Bremen, Yuyi Wang, Juhua Pu, and Ondřej Kuželka. "Fast Algorithms for Relational Marginal Polytopes". In: *International Joint Conference on Artificial Intelligence (**IJCAI**)*. 2021, pp. 4266–4274.

[ILP-21] Jáchym Barvínek, Timothy van Bremen, Yuyi Wang, Filip Železný, and Ondřej Kuželka. "Automatic Conjecturing of P-Recursions Using Lifted Inference". In: *International Conference on Inductive Logic Programming (**ILP**)*. 2021, pp. 17–25.

[KI-20] Timothy van Bremen, Anton Dries, and Jean Christoph Jung. "onto2problog: A Probabilistic Ontology-Mediated Querying System using Probabilistic Logic Programming". In: *Künstliche Intell. (**KI**)* 34.4 (2020), pp. 501–507.

[KR-21] Timothy van Bremen and Ondřej Kuželka. "Lifted Inference with Tree Axioms". In: *International Conference on Principles of Knowledge Representation and Reasoning (**KR**)*. 2021, pp. 599–608.

[PODS-23] Timothy van Bremen and Kuldeep S. Meel. "Probabilistic Query Evaluation: The Combined FPRAS Landscape". In: *ACM Symposium on Principles of Database Systems (**PODS**)*. 2023, pp. 339–347.

[UAI-21] Timothy van Bremen and Ondřej Kuželka. "Faster Lifting for Two-Variable Logic Using Cell Graphs". In: *Conference on Uncertainty in Artificial Intelligence (**UAI**)*. 2021, pp. 1393–1402.