Learning Aggregate Queries Defined by First-Order Logic with Counting

Steffen van Bergerem and Nicole Schweikardt
Humboldt-Universität zu Berlin

How to be a Good Colleague

We want to find a classifier that predicts the popularity of colleagues. For this, via a survey, we gather examples of colleagues and their popularity (see Table 1). In addition, we maintain a list of all cakes that colleagues brought to work, and we model this as a relational structure (see Fig. 1). We note that the following classifier is consistent with the gathered examples: the popularity of a colleague is the number of cakes they brought to work, where chocolate cakes are counted twice. This can also be written as a term in the first-order logic with counting FOC_1 as follows.

Name	Popularity
Alice	5
Bob	1
Carol	2
Dan	3
Emma	1

Table 1: Popularity of colleagues.
(names changed for privacy reasons)

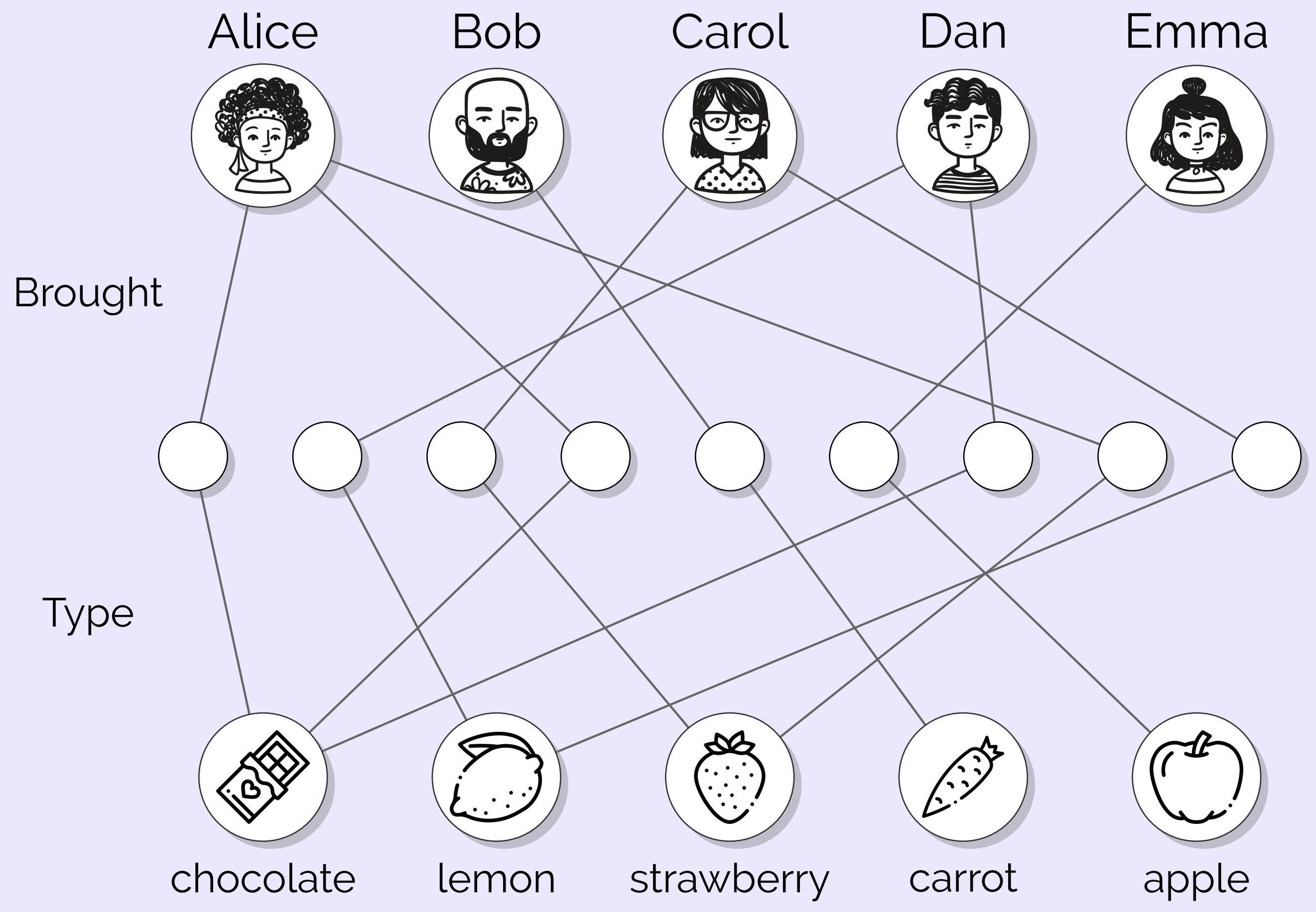

Figure 1: Vertices in the middle represent cakes. Edges indicate the type of the cake and who brought the cake to work.

$$\mathit{popularity}(x) \;=\; 2\cdot\#(c).(\mathit{Brought}(x,c)\wedge \mathit{Type}(c,🍫)) \;+\; \#(c).(\mathit{Brought}(x,c)\wedge\neg\mathit{Type}(c,🍫))$$

Learning from Examples

For fixed $k\in\mathbb{N}_{\geqslant 1}$ and $\ell,q\in\mathbb{N}$, we want to solve the following problem.

Precomputation Given a relational structure $\mathcal{A}$, compute an expansion $\mathcal{A}'$ of $\mathcal{A}$, and compute a lookup table whose size is independent of $\mathcal{A}$.

Given a list of labelled examples of the form $(\bar v,\lambda)\in\left(U(\mathcal{A})\right)^k\times\mathbb{Z}$

Return a term $t(\bar x)\in\mathrm{FOC}_1$ using at most ℓ vertices from $\mathcal{A}$ as constants and having quantifier rank at most q such that $[\![t(\bar v)]\!]^{\mathcal{A}'}=\lambda$ for all given examples $(\bar v,\lambda)$, or reject if there is no such term.

The precomputation step can be seen as building an index structure. After this step, the index structure can be used repeatedly for different lists of labelled examples to compute consistent hypotheses.

After the precomputation step, algorithms are granted only **local access** to the structure. That is, instead of having random access, an algorithm may only access the neighbours of elements that it already holds in memory, initially starting with the elements given in the labelled examples.

Previous works in this framework considered only Boolean-valued concepts. Our **main contribution** is the extension of this framework to integer-valued concepts that are definable in the first-order logic with counting FOC_1.

For **structures of small degree**, that is, structures whose degree is at most polylogarithmic in their size, results of previous works as well as our main contribution can be seen on the right. The running times are measured in the size of the structure.

Grohe and Ritzert, LICS 2017

Boolean-valued concepts definable in first-order logic can be learned in sublinear time.

v. B. and S., CSL 2021

Boolean-valued concepts definable in the first-order logic with counting FOC_1 or the first-order logic with weight aggregation FOWA_1 can be learned in sublinear time after quasi-linear-time precomputation.

Our main result

v. B. and S., ICDT 2025

Integer-valued concepts definable in the first-order logic with counting FOC_1 can be learned in sublinear time after quasi-linear-time precomputation.

FO with Counting

The first-order logic with counting FOC extends FO by counting terms of the form $\#(x_1,\dots,x_k).\varphi$, which can be combined using addition and multiplication. It also adds formulas of the form $P(t_1,\dots,t_m)$, where $P\subseteq\mathbb{Z}^m$ is a predicate and $t_1,\dots,t_m$ are counting terms.

Grohe and Schweikardt (PODS 2018) provided locality results similar to Gaifman normal forms for the fragment FOC_1 of FOC. We heavily rely on these locality results to limit the search space for potential hypotheses.

References

Martin Grohe and Martin Ritzert. *Learning First-Order Definable Concepts over Structures of Small Degree*. LICS 2017.

Martin Grohe and Nicole Schweikardt. *First-Order Query Evaluation with Cardinality Conditions*. PODS 2018.

Steffen van Bergerem and Nicole Schweikardt. *Learning Concepts Described By Weight Aggregation Logic*. CSL 2021.

Steffen van Bergerem and Nicole Schweikardt. *Learning Aggregate Queries Defined by First-Order Logic with Counting*. Accepted at ICDT 2025. Available on https://svbergerem.de.