

# PROBABILISTIC QUERY ANSWERING PARA LENGUAJES MODALES Y DATA CLEANING PARA DATA TREES EN UN MARCO EPISTÉMICO

**Tomás Schitter**

Instituto de Ciencias de la Computación (ICC) - CONICET, Argentina  
tschitter@dc.uba.ar

Los datos estructurados como grafos (*data graphs*) constituyen un modelo ampliamente utilizado para representar relaciones complejas que aparecen de manera natural en varios dominios, como el análisis de redes sociales o la procedencia de datos. Sin embargo, los datos del mundo real suelen ser incompletos, contener errores o presentar inconsistencias con respecto a ciertas propiedades que se esperaría que satisfagan. Estas propiedades pueden expresarse mediante restricciones de integridad en lenguajes lógicos (Navigational Query Languages), como *Regular Path Constraints* o *Reg-GXPath*.

En este trabajo, continuamos el estudio del modelado de la incertidumbre mediante un enfoque epistemológico particular [1,2], en el cual el *data graph* que es observado puede contener errores o inconsistencias. Este *data graph* se considera una versión ruidosa de un grafo limpio que representa correctamente el mundo y que solo se conoce de manera parcial mediante ciertas distribuciones de probabilidad sobre el universo de *data graphs*. Nos centramos en el estudio de dos problemas principales. Por un lado, analizamos la complejidad computacional de distintas variantes del problema de Probabilistic Query Answering (PQA) [3], que en este caso implica dar una definición de qué quiere decir que una fórmula escrita en un lenguaje lógico particular sea válida en este modelo. Vemos que, en la mayoría de los casos, este problema resulta intratable, incluso utilizando lenguajes lógicos modales muy simples. Por otro lado, demostramos que el problema de *Data Cleaning*, el problema de, dado un *data graph* con errores, encontrar el *data graph* limpio más probable, también resulta intratable, incluso si restringimos el problema a una clase muy simple de *data graphs*: los *data trees*. Estos resultados podrían indicar que el modelo es demasiado general y expresivo, y que resulta interesante buscar restricciones naturales del mismo, basadas en datos reales, que mantengan la relevancia práctica del modelo volviendo algunos de estos problemas tratables.

*Trabajo en conjunto con Edwin Pin Baque (Departamento de Matemática, Facultad de Ciencias Exactas y Naturales - UBA) y Sergio Abriola (Instituto de Ciencias de la Computación - CONICET).*

## Referencias

- [1] Abriola, S., Cifuentes, S., Martínez, M.V., Pardal, N., Baque, E.P. (2021). An epistemic approach to model uncertainty in data-graphs. ArXiv, abs/2109.14112.
- [2] Sa, C.D., Ilyas, I.F., Kimelfeld, B., Ré, C., Rekatsinas, T. (2018). A Formal Framework For Probabilistic Unclean Databases. ArXiv, abs/1801.06750.
- [3] Van den Broeck, G., Suciu, D. (2017). Query Processing on Probabilistic Data: A Survey. Found. Trends Databases, 7, 197-341.