The Quest for Schemas in Graph Databases

Extended Abstract

Angela Bonifati
Lyon 1 University & Liris CNRS
Villeurbanne, France
angela.bonifati@univ-lyon1.fr

ABSTRACT

Property graphs are a widespread data model for representing interconnected multi-labeled data enhanced with properties as key/value pairs. These highly expressive graphs are used in a wide range of domains, such as social and transportation networks, biological networks, finance, cybersecurity, logistics and planning, to name a few. Property graphs are the building blocks of future graph ecosystems, in which OLTP and OLAP processes are intertwined with complex advanced processes, such as learning, scientific computing and business intelligence. While property graphs are currently used in a variety of graph databases, a rather fragmented landscape emerges in terms of the supported query and schema languages. In particular, the coverage of schema and constraints is limited if not completely lacking in these systems. In this talk, I will present recent advances in terms of schemas and constraints for property graphs, as part of our work within the LDBC community groups. I will also focus on graph schema discovery and constraint satisfaction following these proposals for property graph schema and constraints. Finally, I will pinpoint future directions of research in this new exciting area of data management.

KEYWORDS

property graphs, graph schemas, graph constraints, schema discovery, Big graph ecosystems

A ROADMAP OF MY TALK

Property graphs are becoming a widely used expressive data model for encoding interconnected data [6]. They are adopted in a wide range of domains, such as social and transportation networks, biological networks, finance, cybersecurity, logistics and planning, to name a few. Property graphs are the building blocks of future graph ecosystems, in which OLTP and OLAP processes are intertwined with complex advanced processes, such as learning, scientific computing and business intelligence. [8]. Property graphs and their key components can be specified without an a priori schema definition. With a schema-less specification, users can create arbitrarily new labeled vertices and edges along with their key-value properties. However, schemas are still useful to provide a description of the underlying data, as well as a guidance during data exploration and query evaluation. Schemas are in this case descriptive and can be extracted from the underlying data by leveraging schema discovery mechanisms. By opposite, graph schemas can be defined prior to populating the graph instance, similarly to relational databases. In this case, schemas are prescriptive and the graph database instance should strictly comply with the specified graph schema.

Descriptive and prescriptive schemas are the two ends of a wide spectrum in which schemas can be an arbitrary mixture of descriptive and prescriptive elements.

The design of graph schema languages. In the talk, I will touch upon early work on the design of a Cypher-like property graph schema language and on the principles behind schema validation and evolution [7]. I will also discuss a recent proposal for PG-Schema, a property graph schema language, that adheres to GQL, the standard query language for property graphs [1].

The design of graph constraints. Database constraints are fundamental building blocks for data quality, normalization and dependency theory.

Property graph constraints are in their early days, and I will discuss our previous work on PG-Keys, property graph key constraints [2], as well as our recent work on cardinality constraints by using a novel class of queries, called threshold queries [3].

Schema discovery from Big Data to Machine Learning. When the schema is left unspecified, schema inference can be applied to graph instances in order to extract the schema information a posteriori. We have studied methods blending graph queries for pre-processing and Big Data approaches (namely, Map Reduce frameworks) in order to discover schema information [7]. An alternative to the above methods, that provides more accurate inference of schemas and allows to holistically consider both properties and labels, uses the Gaussian Mixture Model and a hierarchical clustering algorithms. Variants of schema inference, for the dynamic and incremental cases have been also considered [4, 5].

Looking Ahead. At the end of my talk, I will discuss several open research challenges on the topic of property graph schemas. Schema are part of the data model abstractions of future Big graph ecosystems, as in our vision paper [8], and the quest is certainly not over!

BIOGRAPHY

Angela Bonifati (PhD, 2002) is a Professor of Computer Science at Lyon 1 University and the CNRS Liris research lab, where she leads the Database Group. She is also an Adjunct Professor at the University of Waterloo in Canada since 2020. Her current research interests are on the interplay between relational and graph-oriented data paradigms, particularly query processing, indexing, data integration and learning for both paradigms. She is involved in several grants at Lyon 1 University, including French, EU and industrial grants. She has also co-authored more than 150 publications in top venues of the data management field, and is the recipient of two Best Paper awards (ICDE22, VLDB22 runner up). She has co-authored two books (on Schema Matching and Mapping edited by Springer in 2011 and on Querying Graphs edited by Morgan & Claypool in 2018) and an invited paper in ACM Sigmod Record 2018 on Graph Queries. She was the Program Chair of ACM Sigmod 2022 and she is currently an Associate Editor for both Proceedings of VLDB and IEEE ICDE.

She is an Associate Editor for several journals, including the VLDB Journal and ACM TODS. She is currently the President of the EDBT Executive Board and a member of the Sigmod Executive Committee.

REFERENCES

[1] Renzo Angles, Angela Bonifati, Stefania Dumbrava, George Fletcher, Alastair Green, Jan Hidders, Bei Li, Leonid Libkin, Victor Marsault, Wim Martens, Filip Murlak, Stefan Plantikow, Ognjen Savkovic, Michael Schmidt, Juan Sequeda, Slawek Staworko, Dominik Tomaszuk, Hannes Voigt, Domagoj Vrgoc, Mingxi Wu, and Dusan Zivkovic. 2022. PG-Schema: Schemas for Property Graphs. *CoRR* abs/2211.10962 (2022).

[2] Renzo Angles, Angela Bonifati, Stefania Dumbrava, George Fletcher, Keith W. Hare, Jan Hidders, Victor E. Lee, Bei Li, Leonid Libkin, Wim Martens, Filip Murlak, Josh Perryman, Ognjen Savkovic, Michael Schmidt, Juan F. Sequeda, Slawek Staworko, and Dominik Tomaszuk. 2021. PG-Keys: Keys for Property Graphs. In *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, Guoliang Li, Zhanhuai Li, Stratos Idreos, and Divesh Srivastava (Eds.). ACM, 2423–2436.

[3] Angela Bonifati, Stefania Dumbrava, George Fletcher, Jan Hidders, Matthias Hofer, Wim Martens, Filip Murlak, Joshua Shinavier, Slawek Staworko, and Dominik Tomaszuk. 2022. Threshold Queries in Theory and in the Wild. *Proc. VLDB Endow.* 15, 5 (2022), 1105–1118.

[4] Angela Bonifati, Stefania-Gabriela Dumbrava, Emile Martinez, Fatemeh Ghasemi, Malo Jaffré, Pacome Luton, and Thomas Pickles. 2022. DiscoPG: Property Graph Schema Discovery and Exploration. *Proc. VLDB Endow.* 15, 12 (2022), 3654–3657.

[5] Angela Bonifati, Stefania Dumbrava, and Nicolas Mir. 2022. Hierarchical Clustering for Property Graph Schema Discovery. In *Proceedings of the 25th International Conference on Extending Database Technology, EDBT 2022, Edinburgh, UK, March 29 - April 1, 2022*, Julia Stoyanovich, Jens Teubner, Paolo Guagliardo, Milos Nikolic, Andreas Pieris, Jan Mühlig, Fatma Özcan, Sebastian Schelter, H. V. Jagadish, and Meihui Zhang (Eds.). OpenProceedings.org, 2:449–2:453.

[6] Angela Bonifati, George H. L. Fletcher, Hannes Voigt, and Nikolay Yakovets. 2018. *Querying Graphs.* Morgan & Claypool Publishers.

[7] Hanâ Lbath, Angela Bonifati, and Russ Harmer. 2021. Schema Inference for Property Graphs. In *Proceedings of the 24th International Conference on Extending Database Technology, EDBT 2021, Nicosia, Cyprus, March 23 - 26, 2021*, Yannis Velegrakis, Demetris Zeinalipour-Yazti, Panos K. Chrysanthis, and Francesco Guerra (Eds.). OpenProceedings.org, 499–504.

[8] Sherif Sakr, Angela Bonifati, Hannes Voigt, Alexandru Iosup, Khaled Ammar, Renzo Angles, Walid G. Aref, Marcelo Arenas, Maciej Besta, Peter A. Boncz, Khuzaima Daudjee, Emanuele Della Valle, Stefania Dumbrava, Olaf Hartig, Bernhard Haslhofer, Tim Hegeman, Jan Hidders, Katja Hose, Adriana Iamnitchi, Vasiliki Kalavri, Hugo Kapp, Wim Martens, M. Tamer Özsu, Eric Peukert, Stefan Plantikow, Mohamed Ragab, Matei Ripeanu, Semih Salihoglu, Christian Schulz, Petra Selmer, Juan F. Sequeda, Joshua Shinavier, Gábor Szárnyas, Riccardo Tommasini, Antonino Tumeo, Alexandru Uta, Ana Lucia Varbanescu, Hsiang-Yun Wu, Nikolay Yakovets, Da Yan, and Eiko Yoneki. 2021. The future is big graphs: a community view on graph processing systems. *Commun. ACM* 64, 9 (2021), 62–71.