Knowledge Graph Exploration: where are we and where are we going?

Matteo Lissandrini, Torben Bach Pedersen, Katja Hose
Aalborg University
and
Davide Mottin
Aarhus University

Knowledge graphs (KGs) represent facts in the form of subject-predicate-object triples and are widely used to represent and share knowledge on the Web. Their ability to represent data in complex domains augmented with semantic annotations has attracted the attention of both research and industry. Yet, their widespread adoption in various domains and their generation processes have made the contents of these resources complicated. We speak of *knowledge graph exploration* as of the gradual discovery and understanding of the contents of a large and unfamiliar KG. In this paper, we present an overview of the state-of-the-art approaches for KG exploration. We divide them into three areas: profiling, search, and analysis and we argue that, while KG profiling and KG exploratory search received considerable attention, exploratory KG analytics is still in its infancy. We conclude with an overview of promising future research directions towards the design of more advanced KG exploration techniques.

1. INTRODUCTION

Nowadays, companies like Google, Amazon, and Bosch are using the graph model to represent and store their enterprise knowledge bases [Noy et al. 2019; Schmid et al. 2019]. Moreover, an increasing amount of data is published as RDF datasets and made available as Linked Open Data in different scientific domains [Ghose et al. 2019; Callahan et al. 2013]. We also see widespread adoption of resources like Yago, DBpedia, and WikiData describing entities and facts of general encyclopedic interest, e.g., artists, books, movies, and songs. These networks of rich connections among entities are called *knowledge graphs* (KGs) (see Figure 1 for an example). KGs store highly heterogeneous information and are increasingly adopted to advance more intelligent machine learning systems [Zhou et al. 2020]. As such, they constitute an important source of information for businesses, organizations, and individuals. As a byproduct of their widespread adoption, KGs easily become extremely large and complex. Usually, they are generated (semi-)automatically through the integration of many different (e.g., [Pellissier Tanon et al. 2020; Varga et al. 2016]) and they are updated with unprecedented volumes of data and at unprecedented speed [Le-Phuoc et al. 2011]. Thus, the contents of these KGs have become less and less familiar

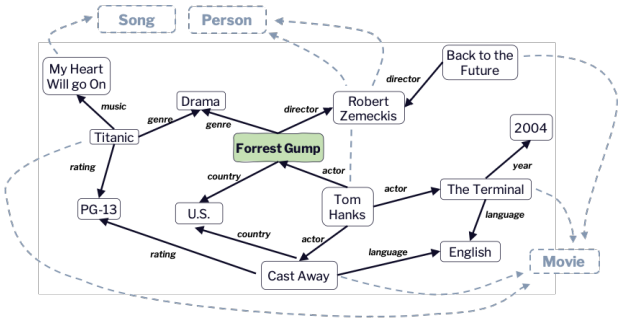

Fig. 1. A simplified fragment of a knowledge graph around the entity "Forrest Gump".

even to domain experts and almost impenetrable to first-time users, calling for exploratory methods on graphs [Mottin and Müller 2017]. In this context, we speak of *knowledge graph exploration* [Lissandrini et al. 2018a] as of the machine-assisted process of progressive analysis of the contents of a KG with the goal of (1) understanding the structure and nature of the dataset at hand, (2) identifying whether the dataset can satisfy the current information need or research question, and (3) retrieving the portion of the dataset that is pertinent to an often vague and hard-to-express information need. These goals are achieved through three main tasks: (i) summarization and profiling, (ii) exploratory search, and (iii) exploratory data analytics.

In recent years, KG exploration in general, and these three areas in particular, have received considerable attention. Hence, we aim at helping researchers and practitioners navigate this scientifically rich area, and identify the most appropriate methods for their business needs and the most promising areas on which to focus future research endeavors. Compared to existing solutions for KG profiling and KG exploratory search, techniques for exploratory KG analytics are still largely unexplored. Thus, we identify promising future directions towards the design of more advanced KG exploration techniques that are able to (a) learn from user interactions to better understand and satisfy a user's information need, (b) efficiently guide the user in their exploration journey towards the most relevant portions of the graph, and (c) enable the exploration of highly heterogeneous datasets.

2. METHODS FOR KNOWLEDGE GRAPH EXPLORATION

Existing approaches in data exploration cover three domains (summarized in Figure 2): (1) methods for KG *profiling and summarization* to distill the most important features and characteristics both of the structure and the contents of a KG; (2) *exploratory search methods* for a gradual discovery and understanding of the items that are pertinent to a vague or underspecified information need; and (3) *techniques for exploratory analytics* to distill salient features from different data subsets.

We organize data exploration methods over a spectrum describing the expertise in domain knowledge, the level of interactivity they expect, and the type of output they are able to produce. Methods that do not require any domain knowledge usually provide high-level information with low granularity and a coarse level of detail. Also, they are typically one-off approaches that do not account for user preferences. On the other hand, methods that require some domain knowledge, e.g., some representative elements of interest or some

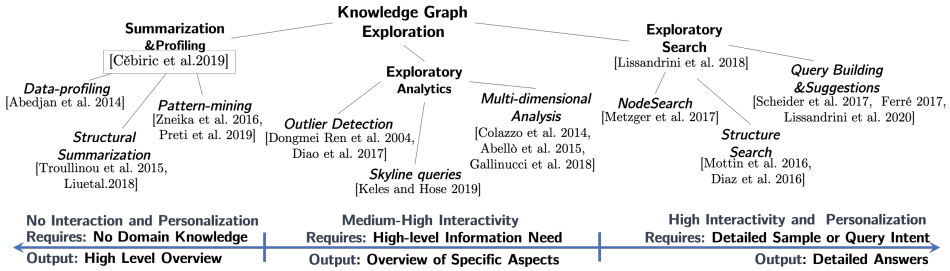

Fig. 2. Taxonomy of KG Exploration techniques and their positioning on the spectrum of features.

initial definition of the query intent, are able to produce more detailed answers with a high level of detail. Moreover, they include user feedback and user interaction to adapt to diverse user needs. Below, we first present summarization and exploratory search as the areas at the extremes of the spectrum. We then describe the special role of exploratory analytics as a middle ground between summarization and exploratory search.

2.1 Profiling & Summarization

Data profiling (e.g., [Abedjan et al. 2014]) is the simplest form of exploration, as it computes basic statistics. For instance, counting the number classes (e.g., Movie) and their instances or summarizing value distributions for specific attributes (e.g., averaging the release year). Their focus is then on frequencies and statistical measures.

Structural summarization [Čebirić et al. 2019; Liu et al. 2018] and *pattern mining* [Zneika et al. 2016; Preti et al. 2019] approaches have been applied to KGs to facilitate understanding the structure of the data as well as to obtain concise representations of the most salient features of their contents. In general, KG summaries either (i) present a compact representation of the main features of the original graph; or (ii) define a new graph derived from the original graph. In the former case, the summary is directly useful for and interpretable by the end-user. In the latter, the summary is used instead of the original graph to provide easier access to answers that are expensive or impractical to compute on the original graph. The main graph summarization techniques extract the schema of the graph [Čebirić et al. 2019], that is a meta-graph composed of high-level patterns and their most representative instances [Troullinou et al. 2015] (e.g., *T. Hanks acted in Forrest Gump*, Figure 1 is a notable instantiation of the high-level –and frequent– pattern *Person acted-in Movie*). Primarily, this is achieved by analyzing frequent substructures, i.e., via *pattern mining* [Preti et al. 2019], or based on node and edge types and their topology [Zneika et al. 2016], e.g., a *Person* can be *actor* or *director* in a *Movie*, but is never playing the role of a *Song*. For instance, we could summarize part of Figure 1 with *Person acted-in and/or directed Movie*.

Overall, *these approaches require no specific domain knowledge and they return a high-level overview of the data.* Thus, they are helpful in the initial exploratory stages since they can assist in evaluating whether a dataset matches the domain of interest, whether any data cleaning is required, and they can help in formulating initial research questions.

2.2 Exploratory Search

While KG summarization allows data understanding by providing a high-level representation of the graph (e.g., a zoomed-out view), exploratory search instead delves into the data itself with the goal of retrieving specific portions of it that are relevant to the current information need (e.g., zooming-in to a subset of items of interest). Yet, contrary to traditional search, where the desired result is well-defined, exploratory search usually starts from a *tentative query* that hopefully leads to answers that are at least partially relevant and that provide cues for the next queries, e.g., inspecting nodes connected to the node *Forrest Gump* to finally identify the relationships with *Back to the Future*.

Hence, exploratory queries change the traditional semantics of the search input: instead of a strict prescription of the desired result set, they provide a hint of what is relevant. This shift in semantics has led to (i) a number of methods following the *search-by-example* paradignm [Lissandrini et al. 2018a] and (ii) methods and interfaces that help the user formalize their intent into a domain-specific query construct that is usually an expansion of the input [Ferré 2017; Lissandrini et al. 2020]. Both have the common goal to overcome one of the main challenges in enabling exploratory search: to avoid complicated declarative languages (e.g., SPARQL) and at the same time retain the flexibility and expressiveness of such languages.

Search-by-example methods receive as input a set of example members of the answer set (e.g., *Tom Hanks* and *Forrest Gump*). The search system then infers the entire answer set based on the given examples and any additional information provided by the underlying database [Mottin et al. 2016] (e.g., other movies by *Tom Hanks*, or a list of *Drama* movies and their actors). This allows retrieving a set of entities similar to some entities of interest [Metzger et al. 2017], or complex structures matching some relevant structure known by the user [Mottin et al. 2016]. *Node and Entity search* allows for automatic completion of a set of seed entities (persons, organizations, places). *Example-based graph search* works similarly to node search but requires a full example (a subgraph or a tuple) to be provided as input. For instance, it is possible to support by-example reverse engineering of (SPARQL) queries from example tuples [Diaz et al. 2016].

To further facilitate the user to formulate a query written in an unfamiliar language and over an unfamiliar dataset, different studies have proposed *query suggestion and refinement* techniques [Lissandrini et al. 2020] and graphical user interfaces [Scheider et al. 2017; Ferré 2017]. Yet, while by-example methods allow for rather vague information needs, query formulation interfaces are designed to help users with a clear information need in writing (relatively simple) queries about specific entities.

Therefore, exploratory search approaches are particularly useful in the later stages of the exploration since they support the user in identifying specific entities, relationships, and structures of interests. They help in answering more fine-grained and specialized information needs but still take into account that the user is not familiar with the dataset. For this reason, particular focus is given to approximate methods [Lissandrini et al. 2018b] and to query suggestion and query refinement techniques.

2.3 Exploratory Analytics

Exploratory Analytics is an iterative, integrated process of data discovery and analytical querying on data which is not well known to the user, e.g., external data. The ability to support analytical workflows for rich KGs has recently received increased attention [Abelló et al. 2015; Colazzo et al. 2014]. The idea is to provide functionalities typical of relational data warehouses, i.e., *multi-dimensional analysis* over knowledge graphs by describing multi-dimensional and statistical within the KG model [Gallinucci et al. 2018; Varga et al. 2016]. This is also motivated by increasing interest from public and private organization to represent business data in specialized knowledge graphs [Schmid et al. 2019]. All these approaches enable a similar approach: to obtain analytical insights on RDF graphs by means of "views" and aggregation operations. For instance, in Figure 1, we could materialize a view that counts for every *Country* and *Genre* the number of *Movies*. Such views are themselves accessible as RDF graphs. Similarly, *skyline queries* are used to find entities that optimize a multi-criteria decision problem to find a set of objects that are of interest to a user because of their *dominance* across multiple attributes [Keles and Hose 2019], e.g., what are the most recent *Movies* with the highest number of *Actors* performing in it.

Finally, *outlier detection* approaches identify elements that are interesting because they are very different from the rest of the elements [Dongmei Ren et al. 2004], e.g., *Actors* who have participated in an unusually high number of *Movies*. One of the most advanced approaches, Dagger [Diao et al. 2017], inspects a triple store and selects different aggregation queries that can describe different entity types based on high-variance values, e.g., whether there is more variability in the number of movies per *Year* or per *Genre* across *Countries*.

In conclusion, exploratory analytics is effective to enable users to identify high-level details w.r.t. facets of the data tailored to specific user needs. In contrast, data summarization approaches are agnostic of the user's information need and only provide a global overview of the data. On the other hand, exploratory search digs into specific data items (entities and relationships) but these searches return very large result sets instead of a more useful aggregate analysis identifying trends and common patterns. Hence, exploratory analytics techniques are a middle ground, where specific summarization methods are applied over large results of an exploratory search. Yet, current approaches usually mimic the same operators proposed for relational data, providing no graph-centric analyses. Moreover, *in these approaches, either the user is required to be familiar with the (complex) query language, or the system is not able to accept any user input to customize the output.* Thus, analytical approaches for KGs are currently missing the ability to reverse engineer analytical queries as well as to suggest appropriate query refinements based on user interactions.

3. FUTURE DIRECTIONS

Analyzing the state of the art (Figure 2), we identify 3 important research avenues for KG exploration: (1) example-based exploratory analytics methods, (2) enhanced interactivity and personalization through machine learning and active learning, and (3) KG exploration applied to the exploration of other datasets, e.g., documents and semantic data lakes.

Example-based exploratory analytics. Among the exploration methods, example-based approaches [Lissandrini et al. 2018a] have the unique advantage to remove the need for

the user to be familiar with the structure of the data and the query language. Yet, to date, example-based approaches exist only for exploratory search tasks and not for exploratory analytics. Exploratory analytics should combine techniques from both summarization and exploratory search: on the one side, similar to the exploratory search case, the user can identify a (usually large) set of elements of interest. Then, these elements are not presented verbatim to the user, instead, data summarization and profiling techniques should be employed to extract context-specific insights. At the core, the *context* and information need have to be inferred from the user-provided examples. Consequently, *example-driven exploratory analytics for KGs* are still mostly unexplored. In particular, they should support example-driven visualizations, summarizations, as well as explanations of results (how is the example related to the user-provided input). Moreover, this should be paired with methods to suggest further analytical explorations during interaction with the user.

Enhanced interactivity and personalization. Data exploration in general, and KG exploration in particular, is a process that cannot be disconnected from the specific user need. The two core tenets are *interactivity* and *personalization*. The two are tightly connected: during interaction with the user, the system can improve and learn more about the user needs to enable personalization. Machine learning and active search [Su et al. 2015] are a promising ground [Milo and Somech 2020] to learn user preferences from interactions and adapt to the user needs. Yet, existing exploratory analytics methods are mainly data-driven: they assume fixed user preferences or refinement criteria based on hardwired rules. The overarching challenge is that, in exploratory methods, we have two unknowns: the information need and the user preference. While some exploratory search methods are able to learn a dynamic notion of *interestingness* from the user input and interaction with the system, e.g., query-reverse-engineering [Diaz et al. 2016] or query suggestion [Lissandrini et al. 2020], other approaches still miss this ability. Similarly, we would expect some form of personalization to also be included in both profiling and exploratory analytics approaches. Personalization is also important to help the user make sense of large result sets since they can provide insights tailored to each individual user. Recent trends in *automatic discovery of insights* naturally complement *automatic visualization recommendation techniques* [Burger et al. 2020; Vartak et al. 2017]. Yet, existing solutions for KGs are still limited to visualizations in the form of 2-dimensions representations of values (e.g., histograms, scatter plots, or heat maps). *For KGs, we need to think about visualizing structures and connections in addition to traditional charts.*

Cross-domain applications. In recent years, KGs have proven highly effective to model heterogeneous data, by mapping entities and concepts that appear in different repositories to equivalent nodes in a KG [Schmid et al. 2019]. This characteristic facilitates data exchange through the integration of different datasets and data models within large and unstructured repositories of data, e.g., *data lakes*, in this case, denoted *semantic data lakes* [Mami et al. 2019]. As such, a KG exploration process is paramount for cross-model and cross-domain exploration workflows. KGs also simplify and represent semantic connections between Web documents [Lissandrini et al. 2015]. Hence, KG exploration techniques could assist the exploration of both Linked Open Data as well as Web documents seamlessly.

ACKNOWLEDGMENTS

Matteo Lissandrini is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 838216.

REFERENCES

ABEDJAN, Z., GRÜTZE, T., JENTZSCH, A., AND NAUMANN, F. 2014. Profiling and mining RDF data with ProLOD++. In *ICDE*. IEEE, 1198–1201.

ABELLÓ, A., ROMERO, O., PEDERSEN, T. B., BERLANGA, R., NEBOT, V., ARAMBURU, M. J., AND SIMITSIS, A. 2015. Using semantic web technologies for exploratory OLAP: a survey. *TKDE 27,* 2, 571–588.

BURGER, I., MANOLESCU, I., PIETRIGA, E., AND SUCHANEK, F. 2020. Toward visual interactive exploration of heterogeneous graphs. In *SEAdata Workshop*. CEUR-WS.org.

CALLAHAN, A., CRUZ-TOLEDO, J., ANSELL, P., AND DUMONTIER, M. 2013. Bio2rdf release 2: improved coverage, interoperability and provenance of life science linked data. In *ESWC*. Springer, 200–212.

ČEBIRIĆ, Š., GOASDOUÉ, F., KONDYLAKIS, H., KOTZINOS, D., MANOLESCU, I., TROULLINOU, G., AND ZNEIKA, M. 2019. Summarizing semantic graphs: a survey. *The VLDB Journal 28,* 3, 295–327.

COLAZZO, D., GOASDOUÉ, F., MANOLESCU, I., AND ROATIŞ, A. 2014. Rdf analytics: lenses over semantic graphs. In *The Web Conference*. ACM, 467–478.

DIAO, Y., MANOLESCU, I., AND SHANG, S. 2017. Dagger: Digging for interesting aggregates in rdf graphs. In *ISWC*. CEUR-WS.org.

DIAZ, G., ARENAS, M., AND BENEDIKT, M. 2016. SPARQLByE: Querying RDF data by example. *PVLDB 9,* 13, 1533–1536.

DONGMEI REN, BAOYING WANG, AND PERRIZO, W. 2004. RDF: a density-based outlier detection method using vertical data representation. In *ICDM*. IEEE, 503–506.

FERRÉ, S. 2017. Sparklis: an expressive query builder for SPARQL endpoints with guidance in natural language. *Semantic Web 8,* 3, 405–418.

GALLINUCCI, E., GOLFARELLI, M., RIZZI, S., ABELLÓ, A., AND ROMERO, O. 2018. Interactive multidimensional modeling of linked data for exploratory olap. *Information Systems 77*, 86–104.

GHOSE, A., HOSE, K., LISSANDRINI, M., AND PEDERSEN, B. W. 2019. An open source dataset and ontology for product footprinting. In *ESWC Satellite Events*. Springer, 75–79.

KAOUDI, Z. AND MANOLESCU, I. 2015. Rdf in the clouds: a survey. *The VLDB Journal 24,* 1, 67–91.

KELES, I. AND HOSE, K. 2019. Skyline queries over knowledge graphs. In *ISWC*. Springer, 293–310.

LE-PHUOC, D., DAO-TRAN, M., PARREIRA, J. X., AND HAUSWIRTH, M. 2011. A native and adaptive approach for unified processing of linked streams and linked data. In *ISWC*. Springer, 370–388.

LISSANDRINI, M., MOTTIN, D., PALPANAS, T., PAPADIMITRIOU, D., AND VELEGRAKIS, Y. 2015. Unleashing the power of information graphs. *SIGMOD Record 43,* 4 (Feb.), 21–26.

LISSANDRINI, M., MOTTIN, D., PALPANAS, T., AND VELEGRAKIS, Y. 2018a. *Data Exploration Using Example-Based Methods*. Morgan & Claypool Publishers.

LISSANDRINI, M., MOTTIN, D., PALPANAS, T., AND VELEGRAKIS, Y. 2018b. Multi-example search in rich information graphs. In *ICDE*. IEEE, 809–820.

LISSANDRINI, M., MOTTIN, D., PALPANAS, T., AND VELEGRAKIS, Y. 2020. Graph-query suggestions for knowledge graph exploration. In *The Web Conference 2020*. ACM, New York, USA, 2549–2555.

LIU, Y., SAFAVI, T., DIGHE, A., AND KOUTRA, D. 2018. Graph summarization methods and applications: A survey. *ACM Computing Surveys 51,* 3, 62.

MAMI, M. N., GRAUX, D., SCERRI, S., JABEEN, H., AUER, S., AND LEHMANN, J. 2019. Squerall: Virtual ontology-based access to heterogeneous and large data sources. In *ISWC*. Springer, 229–245.

METZGER, S., SCHENKEL, R., AND SYDOW, M. 2017. QBEES: query-by-example entity search in semantic knowledge graphs based on maximal aspects, diversity-awareness and relaxation. *Journal of Intelligent Information Systems 49,* 3, 333–366.

MILO, T. AND SOMECH, A. 2020. Automating exploratory data analysis via machine learning: An overview. In *SIGMOD*. ACM, 2617–2622.

MOTTIN, D., LISSANDRINI, M., VELEGRAKIS, Y., AND PALPANAS, T. 2016. Exemplar queries: A new way of searching. *The VLDB Journal 25,* 6, 741–765.

MOTTIN, D. AND MÜLLER, E. 2017. Graph exploration: From users to large graphs. In *SIGMOD*. AMC, 1737–1740.

NOY, N., GAO, Y., JAIN, A., NARAYANAN, A., PATTERSON, A., AND TAYLOR, J. 2019. Industry-scale knowledge graphs: Lessons and challenges. *ACM Queue 17,* 2, 48–75.

PELLISSIER TANON, T., WEIKUM, G., AND SUCHANEK, F. 2020. Yago 4: A reason-able knowledge base. In *The Semantic Web*, A. Harth, S. Kirrane, A.-C. Ngonga Ngomo, H. Paulheim, A. Rula, A. L. Gentile, P. Haase, and M. Cochez, Eds. Springer, 583–596.

PRETI, G., LISSANDRINI, M., MOTTIN, D., AND VELEGRAKIS, Y. 2019. Mining patterns in graphs with multiple weights. *Distributed and Parallel Databases 37*, 1–39.

SCHEIDER, S., DEGBELO, A., LEMMENS, R., VAN ELZAKKER, C., ZIMMERHOF, P., KOSTIC, N., JONES, J., AND BANHATTI, G. 2017. Exploratory querying of SPARQL endpoints in space and time. *Semantic web 8,* 1, 65–86.

SCHMID, S., HENSON, C., AND TRAN, T. 2019. Using knowledge graphs to search an enterprise data lake. In *ESWC*. Springer.

SU, Y., YANG, S., SUN, H., SRIVATSA, M., KASE, S., VANNI, M., AND YAN, X. 2015. Exploiting relevance feedback in knowledge graph search. In *KDD*. ACM, 1135–1144.

TROULLINOU, G., KONDYLAKIS, H., DASKALAKI, E., AND PLEXOUSAKIS, D. 2015. RDF digest: Efficient summarization of RDF/S KBs. In *European Semantic Web Conference*. Springer, 119–134.

VARGA, J., ETCHEVERRY, L., VAISMAN, A. A., ROMERO, O., PEDERSEN, T. B., AND THOMSEN, C. 2016. QB2OLAP: Enabling OLAP on Statistical Linked Open Data. In *ICDE*. IEEE, 1346–1349.

VARGA, J., VAISMAN, A. A., ROMERO, O., ETCHEVERRY, L., PEDERSEN, T. B., AND THOMSEN, C. 2016. Dimensional enrichment of statistical linked open data. *Journal of Web Semantics 40*, 22–51.

VARTAK, M., HUANG, S., SIDDIQUI, T., MADDEN, S., AND PARAMESWARAN, A. 2017. Towards visualization recommendation systems. *SIGMOD Records 45,* 4, 34–39.

WHITE, R. W. AND ROTH, R. A. 2009. *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan & Claypool Publishers.

ZHOU, S., DAI, X., CHEN, H., ZHANG, W., REN, K., TANG, R., HE, X., AND YU, Y. 2020. Interactive recommender system via knowledge graph-enhanced reinforcement learning. *SIGIR 43*, –.

ZNEIKA, M., LUCCHESE, C., VODISLAV, D., AND KOTZINOS, D. 2016. Summarizing linked data rdf graphs using approximate graph pattern mining. In *EDBT 2016*. OpenProceedings.org, 684–685.

Matteo Lissandrini is a postdoctoral researcher at Aalborg University (DK). He holds a Marie Curie fellowship on Example Driven Analytics of Open Knowledge Graphs and he received his PhD in Computer Science from the University of Trento (IT). His research focuses on novel query languages for large scale data mining and information extraction, presenting tutorials on exploration techniques at VLDB, SIGMOD, SIGIR, and ESWC.

Torben Bach Pedersen is a professor at the Center for Data-Intensive Systems (Daisy) at Aalborg University (DK). His research concerns data analytics for "Big Multidimensional Data"-the integration and analysis of large amounts of complex and highly dynamic multidimensional data. He is an ACM Distinguished Scientist, a senior member of the IEEE, and a member of the Danish Academy of Technical Sciences.

Katja Hose is a professor of Computer Science at Aalborg University (DK). Her research is rooted in databases and Semantic Web technologies and spans theory, algorithms, and applications of Data and Web Science incl. knowledge querying, analytics, integration, and sharing. She has been organizing, reviewing, and publishing at top-tier international conferences and journals, e.g., TheWebConf, ISWC, VLDB, SIGMOD, ICDE, VLDBJ.

Davide Mottin is an Assistant Professor at Aarhus University (DK). His research interests include graph mining, novel query paradigms, and interactive methods. He also presented tutorials on exploratory methods for graphs at KDD, VLDB, SIGMOD and coauthored a book on data exploration. He has published relevant works in prestigious venues in data mining (KDD, ICDM), database (VLDB, ICDE), and machine learning (ICLR).