Graph Exploration: Taking the User into the Loop

Davide Mottin Anja Jentzsch Emmanuel Müller

Hasso Plattner Institute
firstname.lastname@hpi.de

Abstract

The increasing interest in social networks, knowledge graphs, protein-interaction, and many other types of networks has raised the question how users can explore such large and complex graph structures easily. Current tools focus on graph management, graph mining, or graph visualization but lack user-driven methods for graph exploration. In many cases graph methods try to scale to the size and complexity of a real network. However, methods miss user requirements such as exploratory graph query processing, intuitive graph explanation, and interactivity in graph exploration. While there is consensus in database and data mining communities on the definition of data exploration practices for relational and semi-structure data, graph exploration practices are still indeterminate.

In this tutorial, we will discuss a set of techniques, which have been developed in the last few years for independent purposes, within a unified graph exploration taxonomy. The tutorial will provide a generalized definition of graph exploration in which the user interacts directly with the system either providing feedback or a partial query. We will discuss common, diverse, and missing properties of graph exploration techniques based on this definition, our taxonomy, and multiple applications for graph exploration. Concluding this discussion we will highlight interesting and relevant challenges for data scientists in graph exploration.

Scope of the Tutorial

The continuously increasing interest in graphs and the growing amount of graph data available on the web require a careful design of data analysis techniques. However, from the user perspective most of the existing techniques appear as a black box that returns results without any explanation. For these reasons our community has resorted to data exploration techniques. In particular, while a huge effort has been devoted to text, relational, and semi-structured data [7], data exploration on graphs (*graph exploration* in short) is still in its infancy. Although many techniques for graphs have been studied in different domains, there is still lack of a unified graph exploration taxonomy. We abstracted user-driven graph exploration properties from techniques proposed in the literature and defined such a unified taxonomy. Our taxonomy consists of three strategies that form the backbone of our presentation along with relevant literature identified so far:

Exploratory Graph Analysis entails the process of casting an incomplete or imperfect pattern query to let the system find the closest match. Such exploratory analysis may return a huge number of results, e.g., structures matching the pattern. Thus, the system is required to provide intelligent support. One such strategy is the well known query-by-example paradigm, in which the user provides the template for the tuples and let the system infer the others.

Refinement of Graph Query Results is needed to deal with the overwhelming amount of results that is typical in subgraph processing. It includes approaches designed to present comprehensive result sets to the user or intermediate results that can be refined further. Instantiations of this kind are graph summaries, top-k methods, query reformulation, and skyline queries.

Focused Graph Mining guides the users to a specific portion of the graph they are interested in. It requires the user to provide feedback in the process to restrict the computation to some portion of the graph. Ego-networks mining belongs to this strategy, since the user search is limited to a particular area of the graph and the algorithms focus on that specific area.

Along a use case based on the large corpus of Linked Data we present the benefits of more user-driven graph exploration methods and tools.

We conclude the tutorial with a number of open research questions, highlighting the huge potential of graph exploration with many challenges still unsolved.

Tutorial Outline

I. **Introduction and motivation** (10 min)

- Necessity of data exploration
- Lack of graph exploration due to the complexity of graph data
- Requirements for graph exploration
- Application for graph exploration systems.

II. **Data Exploration Taxonomy** (20 min)

1. **Exploratory Graph Analysis**: approximate queries and queries-by-example
2. **Refinement of Query Results**:
   query reformulation and refinement, top-k results, skyline queries
3. **Focused Graph Mining**: personalized queries, focused queries, data clustering

III. **User-driven Graph Exploration**

0. **Background** (10 min)
   - Graph models and terminology
   - (Sub)graph isomorphism, graph edit distance
   - Frequent subgraph mining
   - Graph clustering and community detection
1. **Exploratory Graph Analysis** (35 min)
   - Approximate search:
     - Structural preserving: homomorphism [3], strong simulation [12]
     - Incompletely specified patterns [10, 25, 26]
   - By example paradigm:
     - Graph query by example [8, 14]
     - Learning paths [1]
2. **Refinement of Graph Query Results** (35 min)
   - Reformulation and refinement:
     - Graph Query Reformulation with diversity [13]
     - Why-empty and Why-so-many results [22]
     - Result summarization [16, 23]
   - Top-k results:
     - Diversified Top-k Graph Pattern Matching [4]
     - Learning to rank from user-feedback [20]
     - Top-K interesting subgraph discovery in information networks [5, 9]
   - Skyline queries [28, 30]
3. **Focused Graph Mining** (35 min)
   - Focused Graph clustering and outlier detection:
     - Focused clustering providing seed nodes [11, 15, 18]
     - Query-driven outlier detection [6, 29]
   - Space restriction methods:
     - Ego-networks [2] and local community detection [17, 19]
     - Center-piece subgraphs [21]
     - Query-driven graph summarization [27]
   - Reweighting graphs: WIGM [24]

IV. **Real-world use case** (15 min)

- Linked Data graphs: exploration, query refinement, and mining

V. **Open challenges** (20 min)

- Can we *interactively* assist the user toward the retrieval of the correct answer?
- Can we provide *explanations* for the query results?
- Is there a way to efficiently *adapt* the analyses on-demand?
- Can we integrate these techniques into current *graph databases*?

Target Audience

This tutorial is intended for researchers and practitioners interested in big data analytics, graph analytics, and data exploration methods. The tutorial aims at fostering collaborations between several disciplines from CIKM communities, including the database, data mining, information retrieval, and web of data communities. Researchers and students will find interesting ideas and challenges to start research in graph exploration. Moreover, they will get an overview of traditional data exploration forming the basis for graph exploration, as well as a smooth introduction to the basic background concepts in graph mining to understand the core part of the tutorial. For practitioners the tutorial presents a new generation of graph exploration methods, which are applicable and improve on a variety of existing graph databases and data exploration tools.

Related Tutorials

• The tutorial "Overview of Data Exploration Techniques" presented at SIGMOD 2015 by Idreos, Papaemmanouil, and Chaudhuri covers data exploration in relational data, highlighting the importance of data visualization tools, and fast and easy access methods to the data. While the tutorial shows aspect in traditional databases, we focus on graph data and more specifically in new methods designed to involve the user in the process. We will present a selection of the data exploration techniques to introduce preliminary concepts in our tutorial. However, we focus on graph data while their tutorial considers relational and semi-structured data.

• The tutorial "Graph-Based User Behavior Modeling: From Prediction to Fraud Detection" presented at KDD 2015 by Beutel, Akoglu, and Faloutsos covers the important aspect of including user models into graph analysis and shows the potential in fraud detection. Nevertheless, data exploration is not part of the tutorial, since the objective is detecting malicious behaviors through user actions in the system. We focus instead on data exploration methods applied to graphs. Therefore, the content is only marginally related to our proposal.

Presenters

Davide Mottin is a postdoctoral researcher at Hasso Plattner Institute. His research interests include graph mining, novel query paradigms, and interactive methods. He presented graph exploration methods in KDD 2015 and VLDB 2014 and is actively engaged in teaching database, big data analytics, and graph mining for Bachelor and Master courses. He received his PhD in 2015 from the University of Trento.

Anja Jentzsch is a PhD student in the Information Systems Group at Hasso Plattner Institute Potsdam. She is a Linked Data enthusiast, being involved in several Linked Data projects like Wikidata and DBpedia since 2007. She presented a tutorial in graph mining, web of data, and data integration in WWW conference. Currently, she is working on exploring frequent and common graph structures in heterogeneous graph databases.

Emmanuel Müller is professor and head of the Knowledge Discovery and Data Mining group at Hasso Plattner Institute. His research interests include graph mining, stream mining, clustering and outlier mining on graphs, streams, and traditional databases. He presented tutorials in database, data mining, and machine learning conferences such as SDM, ICDM, and ICML. He received his PhD in 2010 from RWTH Aachen University, had been independent group leader at Karlsruhe Institute of Technology (2010 - 2015) and postdoctoral fellow at University of Antwerp (2012 - 2015).

Former tutorials by the presenters

- Emmanuel Müller, Günnemann Stephan, Ines Färber, Thomas Seidl. *Discovering Multiple Clustering Solutions: Grouping Objects in Different Views of the Data*[1]. Tutorial at ICML (2013), ICDE (2012), PAKDD (2012), SDM (2011), ICDM (2010).

[1] Video: `http://bit.ly/1UdV2s7`, Slides: `http://dme.rwth-aachen.de/sites/default/files/public_files/dmcs-icml2013.pdf`

- Knud Möller, Richard Cyganiak, Michael Hausenblas, Jens Lehmann, and Anja Jentzsch. *Realising and Exploiting the EU Data Cloud*[2]. Tutorial at European Data Forum 2012, Copenhagen, Denmark, June 2012.
- Michael Hausenblas, Richard Cyganiak, and Anja Jentzsch. *Practical Cross-Dataset Queries on the Web of Data*[3]. Full-day tutorial at World Wide Web Conference 2012 (WWW2012), Lyon, France, April 2012.

References included in the tutorial

[1] A. Bonifati, R. Ciucanu, and A. Lemay. Learning path queries on graph databases. In *18th International Conference on Extending Database Technology (EDBT)*, 2014.

[2] A. Epasto, S. Lattanzi, V. Mirrokni, I. O. Sebe, A. Taei, and S. Verma. Ego-net community mining applied to friend suggestion. *Proceedings of the VLDB Endowment*, 9(4):324–335, 2015.

[3] W. Fan, J. Li, S. Ma, H. Wang, and Y. Wu. Graph homomorphism revisited for graph matching. *Proceedings of the VLDB Endowment*, 3(1-2):1161–1172, 2010.

[4] W. Fan, X. Wang, and Y. Wu. Diversified top-k graph pattern matching. *Proceedings of the VLDB Endowment*, 6(13):1510–1521, 2013.

[5] M. Gupta, J. Gao, X. Yan, H. Cam, and J. Han. Top-k interesting subgraph discovery in information networks. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 820–831. IEEE, 2014.

[6] M. Gupta, A. Mallya, S. Roy, J. H. Cho, and J. Han. Local learning for mining outlier subgraphs from network datasets. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 73–81, 2014.

[7] S. Idreos, O. Papaemmanouil, and S. Chaudhuri. Overview of data exploration techniques. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 277–281. ACM, 2015.

[8] N. Jayaram, A. Khan, C. Li, X. Yan, and R. Elmasri. Querying knowledge graphs by example entity tuples. *Knowledge and Data Engineering, IEEE Transactions on*, 27(10):2797–2811, Oct 2015.

[9] J. Jin, S. Khemmarat, L. Gao, and J. Luo. Querying web-scale information networks through bounding matching scores. In *Proceedings of the 24th International Conference on World Wide Web*, pages 527–537. International World Wide Web Conferences Steering Committee, 2015.

[10] A. Khan, Y. Wu, C. C. Aggarwal, and X. Yan. Nema: Fast graph search with label similarity. In *Proceedings of the VLDB Endowment*, volume 6, pages 181–192. VLDB Endowment, 2013.

[11] I. M. Kloumann and J. M. Kleinberg. Community membership identification from small seed sets. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1366–1375. ACM, 2014.

[12] S. Ma, Y. Cao, W. Fan, J. Huai, and T. Wo. Strong simulation: Capturing topology in graph pattern matching. *ACM Transactions on Database Systems (TODS)*, 39(1):4, 2014.

[13] D. Mottin, F. Bonchi, and F. Gullo. Graph query reformulation with diversity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 825–834. ACM, 2015.

[14] D. Mottin, M. Lissandrini, Y. Velegrakis, and T. Palpanas. Exemplar queries: Give me an example of what you need. *Proceedings of the VLDB Endowment*, 7(5):365–376, 2014.

[2] `http://2012.data-forum.eu/program/eu-data-cloud.html`

[3] `http://www.wwwconference.org/www2012/program/tutorials/`

[15] B. Perozzi, L. Akoglu, P. Iglesias Sánchez, and E. Müller. Focused clustering and outlier detection in large attributed graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1346–1355. ACM, 2014.

[16] S. Ranu, M. Hoang, and A. Singh. Answering top-k representative queries on graph databases. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1163–1174. ACM, 2014.

[17] Y. Ruan, D. Fuhry, and S. Parthasarathy. Efficient community detection in large networks using content and links. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1089–1098. International World Wide Web Conferences Steering Committee, 2013.

[18] P. I. Sanchez, E. Müller, U. L. Korn, K. Böhm, A. Kappes, T. Hartmann, and D. Wagner. Efficient algorithms for a robust modularity-driven clustering of attributed graphs. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 100–108, 2015.

[19] C. L. Staudt, Y. Marrakchi, and H. Meyerhenke. Detecting communities around seed nodes in complex networks. In *IEEE International Conference on Big Data (Big Data)*, pages 62–69. IEEE, 2014.

[20] Y. Su, S. Yang, H. Sun, M. Srivatsa, S. Kase, M. Vanni, and X. Yan. Exploiting relevance feedback in knowledge graph search. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2015.

[21] H. Tong and C. Faloutsos. Center-piece subgraphs: problem definition and fast solutions. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 404–413. ACM, 2006.

[22] E. Vasilyeva, M. Thiele, C. Bornhövd, and W. Lehner. Answering why empty? and why so many? queries in graph databases. *Journal of Computer and System Sciences*, 82(1):3–22, 2016.

[23] Y. Wu, S. Yang, M. Srivatsa, A. Iyengar, and X. Yan. Summarizing answer graphs induced by keyword queries. *Proceedings of the VLDB Endowment*, 6(14):1774–1785, 2013.

[24] J. Yang, W. Su, S. Li, and M. M. Dalkilic. Wigm: Discovery of subgraph patterns in a large weighted graph. In *SDM*, pages 1083–1094. SIAM, 2012.

[25] S. Yang, Y. Xie, Y. Wu, T. Wu, H. Sun, J. Wu, and X. Yan. Slq: a user-friendly graph querying system. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 893–896. ACM, 2014.

[26] Y. Yuan, G. Wang, L. Chen, and H. Wang. Efficient subgraph similarity search on large probabilistic graph databases. *Proceedings of the VLDB Endowment*, 5(9):800–811, 2012.

[27] N. Zhang, Y. Tian, and J. M. Patel. Discovery-driven graph summarization. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pages 880–891. IEEE, 2010.

[28] W. Zheng, L. Zou, X. Lian, L. Hong, and D. Zhao. Efficient subgraph skyline search over large graphs. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, pages 1529–1538. ACM, 2014.

[29] H. Zhuang, J. Zhang, G. Brova, J. Tang, H. Cam, X. Yan, and J. Han. Mining query-based subnetwork outliers in heterogeneous information networks. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 1127–1132. IEEE, 2014.

[30] L. Zou, L. Chen, M. T. Özsu, and D. Zhao. Dynamic skyline queries in large graphs. In *Database Systems for Advanced Applications*, pages 62–78. Springer, 2010.