In partial fulfillment of the terms for obtaining the PhD degree, Christian Tovgaard Aebeloe will give a lecture on the following subject:

**Decentralized Knowledge Graphs on the Web**

on Tuesday 4th of October 2022, 14:00, in room 0.2.13 at Selma Lagerlöfs Vej 300

**Abstract:**
The increasing popularity of the Web of Data has over the past years caused a significant increase in the number of open knowledge graphs published on the Web. However, the structure of the Semantic Web today relies totally on the data providers to maintain Web services that provide efficient and scalable access to the knowledge graphs, as well as to keep the data up-to-date. Without any monetary incentives to do so, publicly available knowledge graphs often become unavailable and outdated, making it difficult to trust the data available in the knowledge graphs. As a remedy, this thesis investigates and addresses the availability, scalability, and updatability issues with the aim of making knowledge graphs on the Web available, scalable, and updatable.

First, the thesis explores load-balancing in client-server architectures in order to increase the availability and scalability of the knowledge graphs under heavy query processing load by introducing a system that we call WiseKG. To optimize queries in such a setup, WiseKG decomposes the query into star-shaped subqueries and determines, using a cost model that considers factors such as the current load on the server and the data transfer overhead, whether each subquery can be processed more efficiently by the client or the server. A comprehensive experimental evaluation shows that WiseKG significantly improves query processing performance for high demanding workloads compared to state-of-the-art systems while being able to answer more queries without timing out.

Second, the thesis addresses the availability and scalability issues from a different point of view by proposing a decentralized Peer-to-Peer (P2P) architecture for sharing and querying knowledge graphs, called Piqnic. In Piqnic, nodes act as both clients and servers and thus maintain a local datastore (a set of knowledge graph fragments) and a local view over the network (a set of neighboring nodes). Piqnic replicates the fragments over multiple nodes within the network, ensuring the availability of the data even if the uploading node fails. An experimental evaluation shows that by doing so, Piqnic maintains high availability even when a large part of the nodes fails.

Third, the thesis introduces two novel indexing schemes that allow nodes to determine which nodes hold relevant data to which subquery: (1) baseline locational indexes that map each predicate in the query to the nodes that hold relevant data to each predicate, and (2) Prefix-Partitioned Bloom Filter indexes that represent the set of subjects and objects in a fragment as a partitioned bitvector that allows the nodes to ascertain the joinability of two fragments. An experimental evaluation supports the hypothesis that such indexes significantly improve query processing performance while decreasing the network usage compared to Piqnic.

Fourth, the thesis addresses the updatability issue by introducing ColChain, a system that allows the users to collaborate on keeping the data up to date. ColChain divides the P2P network into communities of nodes and relies on community-wide consensus to allow nodes.
to propose and apply consensual updates to the fragments. Furthermore, ColChain represents the entire history of updates to a fragment as a chain of updates, allowing nodes to trace-back faulty updates to their origin, as well as to dynamically roll-back fragments to an earlier version and process queries over them. A comprehensive experimental evaluation shows that ColChain provides efficient community-wide consensual updates to the fragments without incurring a significant cost on query performance.

Fifth, the thesis demonstrates ColChain and introduces a fully functioning ColChain client with a graphical UI. The demonstration highlights how users can navigate the fragments stored by a particular ColChain node. Furthermore, the demonstration shows how users can participate in keeping the fragments up to date, as well as how queries can be processed over a previous version of the fragments.

Last, the thesis introduces the Lothbrok approach to optimizing SPARQL queries in the decentralized setup. In particular, Lothbrok fragments data based on characteristic sets (predicate families) and uses star-shaped query decomposition similar to WiseKG. To accommodate the fragmentation technique, Lothbrok further introduces a novel indexing scheme, called Semantically Partitioned Bloom Filter indexes, that associates the objects in a fragment with the predicates they occur in triples with. Lothbrok nodes use these indexes to build a query execution plan in consideration of cardinality estimations, the compatibility of fragments for the given query, and the locality of the data. Lothbrok is further able to delegate subqueries to other nodes in the network such that the network overhead is minimized. A comprehensive experimental study shows a performance increase of up to two orders of magnitude when comparing Lothbrok to Piqnic and ColChain while the network usage is lowered as well.

Members of the assessment committee are Associate Professor Michele Albano (Chairman), Aalborg University (Denmark), Associate Professor Aidan Hogan, University of Chile (Chile), and Professor Ruben Verborgh, Ghent University (Belgium).

Supervisor for the thesis has been Professor Katja Hose, Aalborg University. Co-supervisor for the thesis has been Associate Professor Gabriela Montoya, Aalborg University.

All interested parties are welcome. After the defense the department will be hosting a small reception in cluster 4.