Semantic description of entities and their relationships represented in the form of a graph, is often referred to as a *Knowledge Graph*. They are the backbone of knowledge enabled applications across several domains such as life science, geoscience, healthcare, Internet of Things, software engineering, etc. Apart from their utility in specific domains, commercial enterprises such as Google, Microsoft, Amazon, LinkedIn, IBM, GE, Siemens, Accenture, etc., use knowledge graphs in their applications. Two Semantic Web technologies, OWL and RDF, which are also W3C standards, play an important role in the representation of a knowledge graph. OWL (Web Ontology Language) is used to build ontologies that act as schema for the data in the knowledge graphs. An ontology represents things, group of things, and relation between them. It is used to formally encode the knowledge in a particular domain. RDF (Resource Description Framework) is used to describe resources in the form of triples (subject, predicate, and object) which in turn form a directed labeled graph with the subject and object as the two nodes and predicate as the edge label. My research work involves working on different aspects of the knowledge graph such as ontology modeling and reasoning, RDF query processing, and developing applications that use knowledge graphs.

Research Background

The projects that I worked on in academia and industry, can be categorized into the following.

- **Distributed OWL Reasoning:** Automated knowledge graph construction is an active area of research where techniques from natural language processing, machine learning, and information extraction are used to build knowledge graphs. There are several well-known large knowledge graphs that have been constructed automatically, such as, DBpedia, YAGO, NELL, and Knowledge Vault. They contain several millions and billions of RDF triples and ontology statements. They keep evolving and growing in size over time. Reasoning is required to infer knowledge and check the consistency of the knowledge graph. It is not possible to reason over such large knowledge graphs on a single machine. In my PhD dissertation, I worked on scalable reasoning over large OWL ontologies. I proposed four strategies to distribute the ontology knowledge across multiple machines and efficiently scale the reasoning process with increasing number of machines in the cluster [1, 2, 3].
- **Scalable RDF Query Processing:** Querying RDF graphs with billions of triples on a single commodity machine is not scalable and would overwhelm the machine. During my internship at Bell Labs, I worked on a scalable RDF query processing approach where the graph is split across multiple machines using a vertex partitioning technique. The triples (subgraph) on each machine are stored in a NoSQL store (MongoDB) to improve scalability. RDF query is analyzed to find the sub-queries that can be run in parallel and translated into a form suitable for MongoDB. This system gave good results when tested with few hundred million triples on a small 3-node cluster [4].
- **Knowledge Graph Applications:** I built several applications in diverse domains such as biomedicine, marketing workflows, social conflicts, oil and gas, that use knowledge graphs and ontologies for background knowledge, data integration, predictive analytics, and content recommendation.

    1. An ontology driven approach was used to integrate biomedical data and manage provenance. This system enabled answering several increasingly complex provenance queries [5].

2. During my summer internship at the Xerox Research Center, I built a temporal consistency checker for marketing workflows. This involved developing a temporal model in OWL and expressing temporal operators as rules. A reasoner was used to provide explanations for temporal inconsistencies.
3. During my summer internship at the IBM T.J Watson Research Center, I worked on an information processing framework for situational awareness of physical events such as protests and marathons [6]. I developed the Protest ontology and the services on top of it such as reasoning, restricted question answering, domain keywords expansion, and data filtering.
4. At GE Research Center, I worked on a predictive analytics application that monitors symptoms and predicts potential problems to reduce the downtime of machines. Knowledge of the machines is captured formally using OWL ontologies and rules are written to predict potential problems. I also worked on a content recommendation application where knowledge graph is used for content filtering and categorization.

Research Agenda

I plan to continue working on areas related to Semantic Web and knowledge graphs as well as their applications.

- **Knowledge Graphs:** Automatic knowledge graph construction and population is an active area of research and it has resulted in several very large knowledge graphs. However, due to its automated nature, several aspects of manually constructed knowledge graphs are either missing or not up to the required standard. I plan to explore these aspects of the knowledge graphs.

  1. Consistency: Adding new triples to a knowledge graph might result in logical inconsistencies. This leads to several interesting questions to investigate. Should we check the consistency of the knowledge graph for each newly added triple? Can approximate consistency check provide the balance between accuracy and performance on very large knowledge graphs? If the addition of triples leads to inconsistency, how do we determine which set of triples to retain and which ones to throw away?
  2. Reasoning: Scalable query time reasoning (backward chaining) techniques have been investigated. Can we further improve the reasoning performance on very large knowledge graphs by making use of the provenance information and constraining the chaining? Information such as the domain and timestamp of statements in a large graph spanning multiple domains and time periods could be made use of.
  3. Multimodal Interaction: Apart from traditional ways of interacting with knowledge graphs such as querying, other modalities of interaction such as voice, augmented, virtual and mixed reality, are gaining traction. There are several interesting aspects to explore such as enabling efficient interaction on large knowledge graphs and privacy issues.

- **Applications:** Semantic Web technologies play a key role in several domains. I am particularly interested in working on applications from the following domains.

  1. Internet of Things (IoT): It is an interconnected network of low power, resource constrained devices that exchange data, provide recommendations and take smart decisions. The tools and

frameworks that support Semantic Web technologies on desktops and server machines are not suitable for resource constrained devices. I plan to take advantage of the latest advances in RDF triple compression and knowledge graph embedding to develop Semantic Web tools for IoT devices.

2. Life Sciences and Healthcare: The domain of life sciences and healthcare, with its complex relations between entities, provide an ideal use case for ontology modeling, data integration, and complex question answering. I would like to engage with the community to get real world data and work on applications that can have a positive impact.
3. ICT4D: In emerging markets such as India, information and communication technologies (ICT) play a crucial role in the development of the underprivileged. I plan to work with the local NGOs in identifying applications that would be useful to the community. Examples of such applications are local language based knowledge graph supported chatbots and voice search.

Due to the interdisciplinary nature of the topics that I am interested in, my research work would involve collaboration with researchers working on machine learning, natural language processing, information extraction, etc., and organizations in domains such as healthcare, pharmacy, and IoT. Through my research, my goal is to not only advance the state-of-the-art but also to make a positive impact on the local communities.

References

[1] Raghava Mutharaju, Pascal Hitzler, Prabhaker Mateti, Freddy Lécué. *Distributed and Scalable OWL EL Reasoning*. ESWC 2015, pp. 88-103.

[2] Raghava Mutharaju. *Very Large Scale OWL Reasoning through Distributed Computation*. ISWC 2012, Part II, pp. 407-414.

[3] Raghava Mutharaju, Frederick Maier, Pascal Hitzler. *A MapReduce Algorithm for EL+*. Description Logics 2010, pp. 464-474.

[4] Raghava Mutharaju, Sherif Sakr, Alessandra Sala, Pascal Hitzler. *D-SPARQ: Distributed, Scalable and Efficient RDF Query Engine*. ISWC 2013, Posters & Demonstrations Track, pp. 261-264.

[5] Satya S. Sahoo, D. Brent Weatherly, Raghava Mutharaju, Pramod Anantharam, Amit Sheth, Rick L. Tarleton. *Ontology-drive Provenance Management in eScience: An Application in Parasite Research*. ODBASE 2009, pp. 992-1009.

[6] Kasthuri Jayarajah, Shuochao Yao, Raghava Mutharaju, Archan Misra, Geeth De Mel, Julie Skipper, Tarek Abdelzaher, and Michael Kolodny. *Social Signal Processing for Real-time Situational Understanding: a Vision and Approach*. SocialSens 2015, pp. 627-632.