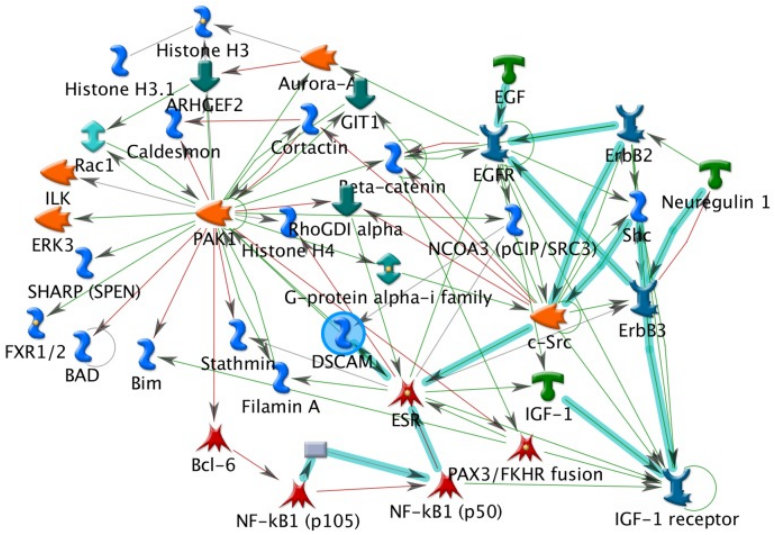
Graph Exploration

Discovery of Hidden Categories in Bio-Medical-Graphs

The world is represented as a graph. Graphs have become an essential data representation in both industry and research for any kind of application scenario ranging from social, financial, retail, logistic, and communication networks in industry up to social networks, biological neural networks, and climate networks in sciences. The reason for the broad success of graphs is the flexibility in representing complex interactions between objects (as edges) as well as additional information (as node and edge attributes). This flexibility is a key enabler for many application domains to exploit the enhanced analytical power of graphs as modern data representations. Graphs from a computer science perspective are very general allowing representation of any data type without explicit need to define the structure of data in advance. In addition, graphs provide from a mathematical point of view a huge theoretical framework from graph theory, network analysis, and complex networks. Such fundamental operations allow graph representations to outperform traditional data representations, which makes graphs a very powerful representation for todays intelligent search, information retrieval, data mining, and in database systems.

Graphs are very challenging. As graphs are very flexible in data representation they are very challenging in data analytics: *Graph Databases* can answer user-driven queries, while on graphs these queries are hard to specify by novice users and fail in many cases. *Graph Mining* has observed high computational cost making automated analysis of most of today's graph sizes in real world applications infeasible. Moreover, graph mining algorithms are exclusively accessible to practitioners and data experts. *Graph Visualization* of large and complex graphs is unable to present the hidden structures that are important for users to understand the complex phenomena. *Graph Exploration* tries to bridge the gap between automated graph mining and manual graph queries or cumbersome interpretation of graph visualization. However, we still lack a fully operational framework in graph exploration that is able to discover the hidden structures of a large graph while being able to interact with novice users.



Our partners in graphs. With Neo4J we have a key player in graph database technology and in open source development of modern graph analytics methods. With the Helmholtz Association as our second partner in this project we aim at graph representations of bio-medical networks.

Project description

Bio-medical data is provided as a large graph by Helmholtz, graph database technology is provided by neo4j, and recent research in graph exploration is provided by HPI. Moreoveor, neo4j provides the expertise, benchmarks, and other dedicated datasets to assess the effectiveness of the developed framework and prototype. Together, we aim to develop novel graph exploration algorithms that can automatically discover hidden facets in a large graph database. Facets are representations of topics, substructures, and relationships and ease the exploration of data proposing succinct summaries to the users. With these facets we will enable faceted graph search, i.e. users will be assisted by facets in queries formulation and graph exploration. In addition, we will enable graph query reformulation within the graph database, i.e. the database will recommend graph queries to the user.

Students have to handle the following tasks during the project:

1. Implementation of graph exploration algorithms
   - Consider graph mining algorithms for automated facet discovery
   - Consider graph query reformulation for recommendation of queries
   - Define novel quality criteria for facet discovery
   - Develop efficient algorithms that prune the search space based on facet quality
   - Consider the evolution and the dynamics of graphs to return real-time results without substantially changing precomputed facets
2. Develop an interactive exploration framework for graph query (re)-formulation, partially integrated into the existing neo4j graph database technology
3. Evaluate the tool on existing large graph data and analyze efficiency and scalability
4. Evaluate the tool w.r.t. its flexibility in handling graph data from other domains
5. Validate the results with domain experts and standard benchmarks
6. Publish research results in scientific publications as well as the prototype as open source

Contacts

Please contact us to apply for the project or ask for more information:

- Coordinator: Prof. Dr. Emmanuel Müller (`emmanuel.mueller@hpi.de`) - Office: E1-01.2
- Mentor: Dr. Davide Mottin (`davide.mottin@hpi.de`) - Office: E1-02.2
- Project partners: Neo4j, Helmholtz Association