Name of Ph.D. Student: **Sajid Yousuf Bhat**
Name of Supervisor: **Dr. Muhammad Abulaish**
Department: **Computer Science, Faculty of Natural Science**
Title: **A Structural Data Mining Framework for Social Network Analysis**

Abstract: Social Network Analysis (SNA) is a multi-disciplinary field dedicated to the modeling and analysis of relations and diffusion processes between various objects in nature and society, and other information/knowledge processing entities with an aim to understand how the behavior of individuals and their interactions translate into large-scale social phenomenon. Due to exploding popularity of online social networks and availability of huge amount of user-generated contents, there is a great opportunity to analyze social networks and their dynamics at different resolutions and levels of granularity. This has resulted in a significant increase in research literature at the intersection of Computing and Social Sciences, leading to several techniques for social network modeling and analysis in the area of Machine Learning and Data Mining.

In this thesis, we explore some of the current challenges related to the analysis of large-scale social networks, which includes community analysis, link prediction, and information diffusion. We present a comprehensive survey of the various data mining techniques proposed in literature to address these challenges. We present a social network analysis framework which mainly aims to address the problem of community analysis, including overlapping community detection, community evolution tracking, and hierarchical community structure identification in a unified manner. Communities in social networks are considered important because they can often be closely related to the functional units of a system, e.g., groups of individuals interacting with each other in a society, web pages related to similar topics, criminal gangs in social networks, and topic-centered discussion groups. Community detection from social networks is highly challenging and depends on various factors including, whether the definition of community relies on global or local network properties, whether overlapping communities are allowed, whether link weights are utilized, whether outliers are considered, whether hierarchical nature of communities needs to be considered, and whether communities need to be monitored and tracked over time in dynamic social networks.

We present a density-based formulation for overlapping community detection which incorporates a novel distance function utilizing link weights of the interaction graph (if available) of a given network, besides able to find communities in unweighted networks. Other novelty of the proposed density-based formulation is that it does not require any neighborhood threshold (ε) to be specified by the users, which is mostly difficult to determine for traditional density-based community detection methods. Instead, it automatically determines a local version of ε for each node locally from the underlying network. The approach is validated against some of the recently proposed state-of-the-art community detection methods on real-world and synthetic datasets in terms of both efficiency and significance of the identified communities.

The proposed density-based formulation is also extended for tracking the evolution of overlapping community structures in dynamic social networks using an adaptive approach. Unlike most of the existing dynamic community detection methods, instead of re-processing all nodes for every new time-step, the proposed method adapts a known community structure (identified at a previous time-step) to the changes occurring in a network by re-processing only a set of active nodes for a new time-step. Moreover, the

proposed method uses an efficient log-based approach to map evolutionary relations between the communities identified from two consecutive network states of a dynamic social network. This approach considerably reduces the number of community comparisons and enhances the efficiency of the proposed method in comparison to the traditional community tracking methods where each pair of communities identified at consecutive network states needs to be compared. The proposed method identifies all evolutionary events like *birth*, *death*, *merge*, *split*, *growth*, and *shrinkage* of communities, and it does not need an ageing function to remove old interactions to be able to identify these events. Experimental results on some dynamic network benchmarks show that the proposed method significantly identifies and tracks communities in dynamic networks.

The overlapping community detection and tracking method is also generalized to identify hierarchical structures of communities in social network. The problem of identifying two consecutive hierarchical levels is modeled as mapping community evolution between community structures identified the same state of the underlying network at two different values of an input parameter (η). We also present a heuristic approach for estimating an optimal value of η, based on maximizing the modularity score of the identified community structures.

Out of various community detection methods proposed in literature, a very few present the utilization of the identified communities. In this regard, we present a real-world application of the identified community structures in online social networks. We propose various community-based features to learn classification models for spammer detection in online social networks. Finally, we present a spam prevention approach which involves restricted legitimate communication only between the users within high-level communities.