Final Seminar Report for

**CPSC 7192, Graduate Seminar**

**Fall 2017**

Student Name: Xiangwen Liu

Student ID: T00610264

Integrated computing

Of Information Science

4th Dec 2017

**Title: New Developments in Mining Big Data**

**Speaker:  Dr. Xiaowei Xu**

1) Motivation and research focus

Many algorithms find clusters by maximizing the number of intra-cluster edges. While such algorithms find useful and interesting structures, they tend to fail to identify and isolate two kinds of vertices that play special roles – vertices that bridge clusters (hubs) and vertices that are marginally connected to clusters (outliers).

In the area of data mining, machine learning, bioinformatics, database management systems and high-performance computing. Big data is highly complex with many interrelated elements, which can be modeled as a network by connecting the dots (i.e. the interrelated elements). The network provides us a big picture of the underlying complexity, but it is often too big to discover any useful hidden patterns from millions of nodes and links.

2) Significance of the issue investigated

Identifying hubs is useful for applications such as viral marketing and epidemiology since hubs are responsible for spreading ideas or disease. In contrast, outliers have little or no influence, and may be isolated as noise in the data.

3) What kind of approach is taken?

They proposed a novel algorithm called SCAN (Structural Clustering Algorithm for Networks), which detects clusters, hubs and outliers in networks. It clusters vertices based on a structural similarity measure. The algorithm is fast and efficient, visiting each vertex only once. An empirical evaluation of the method using both synthetic and real datasets demonstrates superior performance over other methods such as the modularity-based algorithms.

4) The outcomes.

They applied SCAN to some real world networks including finding conferences using only the NCCA College Football schedule, grouping political books based on co-purchasing information, and customer data integration. In addition, they compared SCAN with the fast modularity-based algorithm in terms of both efficiency and effectiveness. The theoretical analysis and empirical evaluation demonstrate superior performance over the modularity based network clustering algorithms.

**Title: Android Security: Attacks, Defenses, and Education**

**Speaker:  Dr. Wenliang Du, Professor at Syracuse University**

1. Motivation and the research focus

Some of the recent attacks and vulnerabilities have been identified including the code injection attacks on the HTML5-based apps and our investigation work on understanding how vendor customization can lead to security problems in Android.

2) Significance of the issue investigated

With this technology, even if the operating system on the phone is compromised, user’s data and actions can still be protected. How to improve the security of the Android operating system.

3) What kind of approach is taken?

They built the system using ARM’s TrustZone technology and developed a novel design that integrates TrustZone with the Android OS, allowing normal apps to leverage the TrustZone functionalities.

4) The outcomes.

They build a prototype phone using the TrustZone-enabled HiKey board to evaluate our design.  They also developed many hands-on lab exercises for security education, including two labs on Android security.

**Speaker:  Dr. Mathias Brochhausen**

**Title:  Improving Data Quality in Biomedical Data using Semantic Web Technologies**

1. Motivation and the research focus

To analyze data and link new data from heterogeneous data sources, and how to overcome the problem of data silos is urgent to infer implicit knowledge from data.

2) Significance of the issue investigated

Using semantic web technologies allows using formal logic to infer implicit knowledge from data. They Semantic Web Technologies, especially ontologies in medical knowledge management systems.

3) What kind of approach is taken?

Using Semantic Web Technologies (SWT) to build features that correspond with data quality dimensions (e.g. uniqueness). Using logical inference in SWT and help overcome problems regarding completeness and consistency. They also created a knowledge representation approach based on ontological realism fosters accuracy.

4) The outcomes.

They improved quality and consistency of evidence type determination by using axiomatic definitions and enabling pulling all data that is about blood pressure using one concept and enabling the representation of what makes, for instance “BP STANDING” different from “BP LYING”, using semantic web can influence a number of data quality dimensions. They developed and co-developed multiple ontologies coded in Web Ontology Language (OWL) such as the ACGT Master Ontology, Ontology for Biobanking, the Drug Ontology, and the Ontology of Document Acts.