Social Network Analysis as Knowledge Discovery process: a case study on Digital Bibliography

Michele Coscia\textsuperscript{1}, Fosca Giannotti\textsuperscript{2} and Ruggero G. Pensa\textsuperscript{3}

\textsuperscript{1} Dept. of Computer Science University of Pisa
\textsuperscript{2} KDD-LAB ISTI CNR Pisa
\textsuperscript{3} Dept. of Computer Science University of Turin
Introduction

Interest in Social Networks:

- Statistical Analysis and Models
- Data Mining: Community Discovery and Temporal Evolution
- Graph Mining and Data Mining Models

Digital Bibliographies (NCBI, DBLP, ...):

- A powerful instrument that collects a great amount of data about scientific publications
- Starting from these data it is possible to construct a co-authorship network
- Enables to model the underlying collaboration links among different researchers
Traditional Social Network Analysis

Given a particular phenomenon we represent it with a graph and then we analyze some key features:

- Vertex Degree
- (Giant) Component
- Clustering and 2-Neighborhood Clustering

Limits of Social Network Analysis

- Static Analysis
- Poor Semantic
- There is no “standard” approach
If we apply Data Mining techniques we can:

- Provide dynamic analysis
- Mine knowledge in networks
- Create a common framework for the analysis

In practice define for Social Network Analysis a Knowledge Discovery Process made of:

- Data Selection
- Preprocessing
- Data Transformation
- Analysis
- Interpretation
First Implementation: Bibliography Analysis

- **Digital Bibliographies**
- **Co-Author Warehouse**
- **Keywords**
- **Authors**
- **Author Class Index**
- **Domain Expert**
- **Query**
- **Co-Author Network**
- **Statistical Parameters**
  - (Global Analysis)
- **Graph Patterns**
- **Mining Parameters**
  - (Local Analysis)

**Data Selection**
**Preprocessing**
**Data Transformation**
**Analysis**
**Interpretation**

**Knowledge**
Case Study

54 = \{\text{Bayesian, At Markov}\}
63 = \{\text{Novel, Opt in Comput, In}\}
104 = \{\text{Temporal, Artificial}\}
Case Study: Conference Dominant Class

Objective: Find what class for a given conference is the most important

Mining Approach
The class relevance is the number of graphs with at least one author belonging to it on the number of total graphs in the graph dataset

Statistical Approach
A class is important if there are many authors belonging to it and these authors have a high local cluster value.

The two levels of analysis confirm each other and mining techniques can be used to support the statistical analysis.
Conclusion and Future Work

• We have introduced a framework for bibliography analysis
• We have showed the potential impact of combining global analysis tasks and local pattern mining approaches

• It might be possible to define some global analysis that enable a direct confrontation of different research groups
• The mining process should work not only on classes of research domains, but also on the single researcher
• Embed our approach in a graph OLAP framework