Analysis of Complex Networks

MAP-I Curricular Unit in Technology

Summary

This document presents a proposal of a Ph.D. level course intended as a Curricular Unit in Technology for the MAP-I Doctoral Programme in Informatics. It is offered jointly by (i) Universidade do Porto, Departamento de Ciência de Computadores and (ii) Universidade de Aveiro, Departamento de Electrónica, Telecomunicações e Informática. The course will cover complex networks concepts, theories, visualizations and algorithms. Topics include methods for structural characterization, pattern detection, information extraction, influence detection, and business intelligence in social media.

Instructors

Fernando Silva	DCC, FCUP	fds@dcc.fc.up.pt
José Luis Oliveira (Coordinator)	DETI, UA	jlo@ua.pt
Srgio Nunes	DEI, FEUP	sergio.nunes@fe.up.pt

Course Objectives

A wide variety of real life structures can be intuitively represented by complex networks, which have received increased attention in recent years. Mining interesting features from these networks is a very important task with an inherent multidisciplinary impact, and practitioners have already available a large set of measurements. From the applicational point of view, this course will mainly focus on biological and social networks, but the methodologies given are general and flexible enough to be applied to any kind of network. The first part of the course covers the fundamentals in graph theory and network metrics, visualization and models. The second part follows a data mining approach, with emphasis on subgraph mining, network comparison and the challenges of very large scale network analysis. The third part addresses the fundamentals of web information retrieval and link analysis. The final part of the course presents several social and information network mining applications.

- "Networks and their Applications", Cesar Hidalgo, MIT
- "Network Theory", Mark Newman, University of Michigan
- "Networks", Jon Kleinberg, Cornell University
- "Information Networks", Amin Saberi, Stanford University
- "Analysis of Biological Networks", John Essigmann, MIT
- "Biological Networks", Dana Peer, Columbia University
- "Introduction to Bioinformatics", Natasa Przulj, Imperial College London

Learning Outcomes

At the end of the this course the students should be able to:

- explain key concepts and algorithms in complex network analysis;
- apply a range of techniques for characterizing network structure;
- define methodologies for analyzing networks of different fields;
- demonstrate knowledge of recent research in the area and exhibit technical writing and presentation skills.

Teaching Methods and Evaluation

The course is composed by theoretical-practical classes, discussions, student assignments and presentations. The student evaluation is based on the following key components:

- 50% Final Exam
- 40% Course Assignments and Presentations
- 10% Participation in Class

Course Content

The course is structured in four parts: (I) Introduction to Network Analysis, (II) Graph Mining and Patterns Discovery, (III) Information Networks and the World Wide Web, (IV) Social and Information Networks: Techniques and Applications.

Part I - Introduction to Complex Network Analysis

1. **Graph Theory Concepts:** basic definitions; graph types; graph examples; graph representations, classical graph algorithms.
2. **Network metrics:** degree, clustering coefficient, diameter, density, shortest paths, centrality, communities.
3. **Network visualization:** graph formats; graph drawing; graph layout methods and algorithms; software experimentation.
4. **Graph models:** Erdös-Rényi random models; small-world and Watts-Strogatz model; scale-free networks, preferential attachment and Barabsi and Barabási-Albert model; other models (ex: hierarchical and geometric).

Part II - Graph Mining and Patterns Discovery

1. **Subgraph Mining:** subgraphs as local metrics; subgraph census; network motifs concept and algorithms; graphlet degree distribution; frequent subgraph discovery methods.
2. **Network comparison:** comparative approaches; structural distance metrics; network alignment; network clustering; graph querying.
3. **Large Scale Network Analysis:** peta-scale networks; graph storage; scalable algorithmic design patterns; MapReduce approaches; Pegasus system and graph mining as a matrix multiplication problem; Pregel computation models and primitives.

Part III - Information Networks and the World Wide Web

1. **Information retrieval overview:** Scoring, TF-IDF, Vector space model.
2. **Web crawling and link analysis:** The problem of ranking, PageRank, Hubs and Authorities.
3. **Information extraction:** Named entity recognition, normalization, relation extraction.

Part IV - Social and Information Networks: Techniques and Applications

1. **Social web mining case studies:** Analysis of explicit and implicit user interaction networks, semantic networks, folksonomies.
2. **Business intelligence:** Information extraction and sentiment analysis of social media streams.
3. **Influence detection and expert finding:** Measures of user influence. Identification of user roles and topic experts in online communities.

Main Bibliography

- Newman, M. "Networks: An Introduction", Oxford University Press, 2010.
- Junker, B. and Schreiber, F. "Analysis of Biological Networks". Wiley, 2008.
- Easley, D., Kleinberg, J. "Networks, Crowds, and Markets: Reasoning About a Highly Connected World". Cambridge University Press, 2010.
- Manning, C., Raghaven, P. and Schutze, H. "Introduction to Information Retrieval", Cambridge University Press, 2008.

Additional Bibliography

- West, D. "Introduction to Graph Theory (2nd Edition)". Prentice Hall, 2001.
- Newman, M., Barabsi, A. and Watts, D.. "The Structure and Dynamics of Networks". Princeton University Press, 2006.
- Wasserman, S. & Faust, K. "Social network analysis: Methods and applications". New York: Cambridge University Press, 1994.
- Russel, M.A. "Mining the Social Web". O'Reilly Media, Inc. 2011.

Software

- Cytoscape: An Open Source Platform for Complex Network Analysis and Visualization (multiplatform/java based) - http://www.cytoscape.org/
- Gephi: an open source graph visualization and manipulation software (multiplatform/java based) - http://gephi.org/
- Pajek: a program for large network analysis (Windows) - http://vlado.fmf.uni-lj.si/pub/networks/pajek/
- GraphCrunch 2: Software tool for network modeling, alignment and clustering (Windows and Linux) - http://bio-nets.doc.ic.ac.uk/graphcrunch2/
- FANMOD: a tool for fast network motif detection (Windows, Linux, Mac) - http://theinf1.informatik.uni-jena.de/ wernicke/motifs/
- SNAP: a C++ libary for working with massive network datsets (Windows, Linux, Mac) - http://snap.stanford.edu/

Instructors Team

The team of instructors is actively involved in research in the field of complex network analysis and information retrieval. A short biography of the instructors is provided below:

Fernando Silva (`http://www.dcc.fc.up.pt/~fds`) is an Associate Professor at the University of Porto, Computer Science Department. He holds a PhD from the University of Manchester, UK, and the Habilitation in Informatics from the New University of Lisbon. He is the founding Director of CRACS (Center for Research in Advanced Computing Systems – `http://cracs.fc.up.pt`). His main research interests are in parallel and distributed computing, scalable algorithms, complex network analysis and programming languages. He has been involved in more than 20 research projects and has more than 100 indexed publications.

José Luis Oliveira (`http://www.ieeta.pt/~jlo`) holds a PhD from University of Aveiro where is currently an Associated Professor. He is the coordinator of the bioinformatics group of University of Aveiro (http://bioinformatics.ua.pt/) and his main research interests are in the area of distributed systems, information retrieval and biomedical informatics. He has been involved in more than 20 research projects and he has more than 200 publications in books, book chapters, journals and conferences.

Srgio Nunes (`http://www.fe.up.pt/~ssn`) is an Assistant Professor at the Faculty of Engineering of the University of Porto (FEUP), in Portugal, and a Senior Researcher at the Information and Computer Graphics Systems Unit at INESC TEC. He has a Masters in Information Management (FEUP, 2004) and a PhD in Informatics Engineering (FEUP, 2010) in the field of Information Retrieval, with a dissertation entitled "Information Retrieval on Time-Dependent Collections". His main research interests are in the fields of Information Retrieval, Information Management, Web Technologies, and Multimedia.

We expect to have the contribution of **Pedro Ribeiro** (`http://www.dcc.fc.up.pt/~pribeiro`) from DCC-FCUP, a spcialist in scalable algorithm for motif discovery in complex networks.