INNOVATIVE APPROACHES TO COMMUNITY DETECTION

Marco Buzzanca

marco.buzzanca@dieei.unict.it

Introduction

One of the most attractive problems in network science deals with the **identification of mesoscale structures** of a complex network, a topic of intensive research activity across multiple disciplines (a). Its importance relies in the ability to unveiling communities of units that, in turn, can be used to **explain some hidden behaviors of networks** emerging as the result of the complex interaction patterns among nodes or entities.

Community detection has also been successfully used to analyze the structure of networks and for **modeling several kinds of interactions**, such as social relationships, genetic interactions among biological molecules or trade among countries.

But what is it really a community? Intuitively, it could be defined as a set of entities that are "**close**" to others, where the concept of closeness is based on a similarity measure and is usually defined over a set of entities. In practice, a network is said to have a community structure if its nodes can be grepped together in set of **nodes that are densely connected**. This latter definition is one of the most acknowledged in literature, although a complete agreement about the definition is still not reached, as there are many cases where communities may be better represented with **different definitions**.

During my time as a PhD, **community detection** has always been a **central topic** of my activity as a student. It is only natural that my final year would take advantage of the accumulated knowledge to try and take on the problem with **a different approach**.

Research Topics

My research consisted in the **exploration of different fields** which knowledge can be used in order to **create solutions** to the community detection problem.

One of the fields that I explored was **information theory**. There are measures in information theory that can deduce the amount of **information contained in a group of nodes**. These measures can be exploited to **assess the quality of a partition** in communities. Concepts from information theory have been used before in community detection (b), but, for the most part, **existing literature is limited to single-layered networks**, hence I researched information-theoretic approaches to community detections in **multiplex networks**.

Another field I have explored is **evolutionary algorithms**. Evolutionary algorithms are used to find **solutions to problems that are intractable** in terms of traditional approaches. The advantage of evolutionary algorithms is that they do not assume or enforce a model on the dataset, so they are **free from modeling bias**. This fact make them **very attractive** if applied to community detection, due to the lack of the agreement on the very definition of community.

Research Collaboration

Universitat Rovira i Virgili

Tarragona, Spain, November 2016-March 2017

During the time I spent in Tarragona, I collaborated with professor Manlio De Domenico on the topic of **community detection in multiplex networks**. I have studied **citation networks**, expanded my knowledge of **statistics**, and researched information theory.

In particular, I have extracted the **author/collaboration network** out of the sciMAG dataset (c), and contributed to the preprocessing step in order to attempt **disambiguation**.

I also have attended lessons about **Bayesian inference** that provided me statistical tools and a better understanding of how to approach statistical data.

At last, I devised and implemented an **information-centric framework** which uses two measures from information theory to **fine-tune the relax rate**, an important parameter of the multiplex MapEquation algorithm.

Publications

Evaluating the community partition quality of a network with a genetic programming approach

5th Intl. Workshop on Complex Networks and their Applications

Milan, Italy, November 20-December 2, 2016

This workshop gathers researchers and enthusiasts in the field of **complex networks** from around the world. I presented an **article** describing in detail a **genetic programming framework** that is able to infer the quality function of a partition in communities. This article provides the **reasoning behind the choices made** when designing the framework.

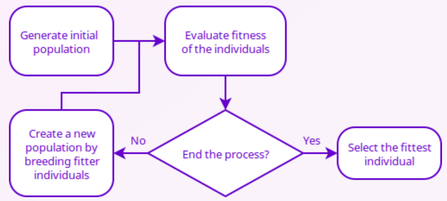
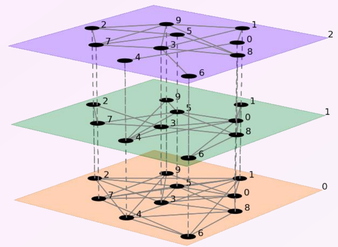


Fig. 1 The process behind genetic programming

Optimal Community Detection in Non-Interconnected Multiplex Networks

Algorithms

Algorithms is an open access journal which provides an advanced forum for **studies related to algorithms** and their applications. The **article** I wrote introduces **two information theory measures** that are used for parameter inference in combination with the Multiplex MapEquation algorithm.



The parameter that is to be inferred in Multiplex MapEquation is the **relax rate**, which represents the degree of coupling between two layers. The article is ready for submission.

Future Works

Even though my days as a PhD have ended, there are still a few **unfinished projects** that I plan to bring to completion. The problem of full **disambiguation** for the sciMAG dataset is still not solved, even though there is significant progress made in the **institution disambiguation**. Moreover, I plan to improve the **genetic programming framework** I devised in order to reduce the running time and modify the fitness function so that it better suits the quality function search problem.

Literature cited

(a) S. Fortunato, 2009. *Community detection in graphs.* Physics Reports, Volume 486, Issue 3-5, p. 75-174.

(b) M. Rosvall, D. Axelsson, C. T. Bergstrom, 2009. *The Map Equation.* The European Physical Journal Special Topics.

(c) M. De Domenico, E. Omodei, A. Arenas, 2016. *Quantifying the Diaspora of Knowledge in the Last Century.* Applied Network Science 1, 15.