Graph Theory and Galaxy Evolution

SPICA 2019: Exploring the Infrared Universe: The Promise of SPICA

Orestis Pavlou, Andreas Efstathiou and Vicky Papadopoulou Lesta
Astrophysics and High Performance Computing Research Group
European University Cyprus
http://ahpc.euc.ac.cy

Graph Theory

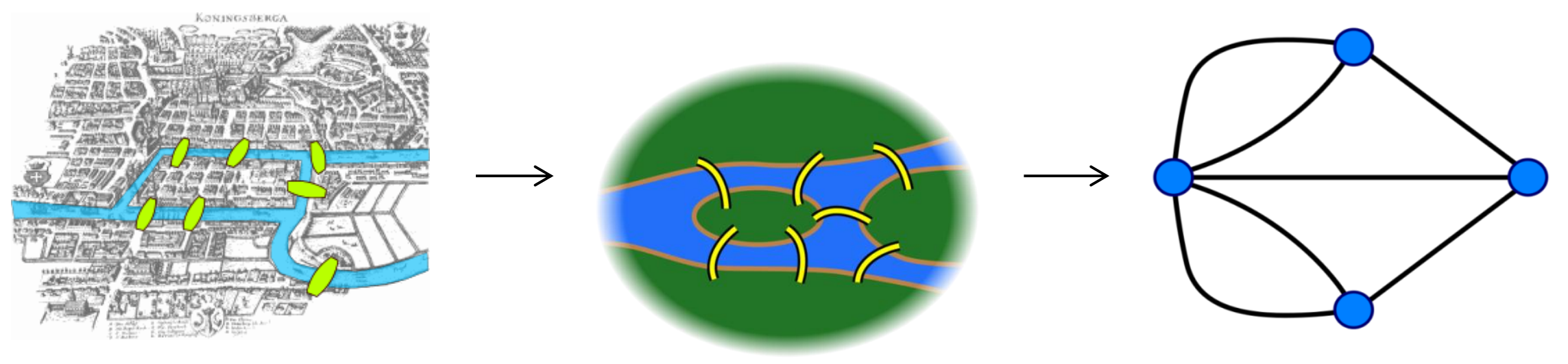
Introduction

The foundations of Graph Theory were laid down by Leonhard Euler in 1736 in his attempt to solve the seven bridges of Konigsberg problem. Since then, Graph Theory has been successfully applied for modeling and solving problems in various sciences. This can be attributed to its interdisciplinary applicability in diverse sciences, spanning from Social and Political Sciences, Biology, Chemistry to Neuroscience [Bullmore and Sporns 2009] and Astrophysics [Farrah et al. 2009], but also due to the rich and powerful knowledge that has been developed related to Graph Theory over the centuries. The power of Graph Theory is also demonstrated through the fact that in new academic areas, which in recent years accomplished remarkable developments, such as Network Science and Graph Databases, Graph Theoretic notions constitute a central theoretical framework and tool.

A graph, in its basic version, consists of a simple but concrete modeling for describing pairwise relationships (edges of the graph) between a set of entities (nodes of the graph). The *multiple* possible *representations* of a graph, *topologically* and as a *matrix*, allow for the deployment of several strong scientific tools for their analysis, resulting to a rich development of Graph Theoretical results, spanning from structural and functional properties that characterize the corresponding systems (networks) to efficient computational solutions for addressing important problems on such systems.

Example of a real problem and its modelling as a graph:

Konigsberg's bridges problem:



Graph Properties

- Node degree: the number of edges incident to the vertex
- Node neighborhood: the nodes directly connected to a node
- Path: a finite or infinite sequence of edges which joins a sequence of vertices
- Cycle: a path where the first and last nodes are equal
- Distance: the length (number of edges) of the shortest path between two nodes

Clustering

Clustering is a fundamental task for the exploration and analysis of big data sets. It is the task of partitioning a set of objects into groups (clusters) satisfying some criteria and optimizing an objective function. Common grouping criteria are the 'similarity' between members of a group or their high interconnection or dependency. Common optimization objectives are the size of the clusters to be equal or balanced, or the interconnection/dependency between different clusters to be minimum.

Clustering techniques [Bader et al 2012] are divided in two categories: (i) hierarchical: employ a multilevel approach for finding the clusters. (ii) partitioning: separating main graph into smaller subgraphs based on their properties.

Known techniques for clustering problems include:

- ✓ Probability theory (Markov Chains)
- ✓ Spectral Graph Theory (using eigenvectors of the corresponding similarity graph)
- ✓ k-means algorithms (determines all clusters at once)
- ✓ Greedy Agglomeration (starting with a large number of small cluster and iteratively merge to larger ones)
- ✓ Minimum Cut Trees
- ✓ Dynamic Clustering (considering changes of the network)
- ✓ Local search and evolutionary algorithms [Bader et al. 2012]

Applications

Applications of clustering range from Statistics, Scientific Computing, Computer Science (Data Mining and machine learning, parallel computing, web searching), Astrophysics (galaxy evolution), Biology and life sciences, (gene analysis, phylogeny, genomics, and proteomics [Zhao and Karypis 2005] and Physiology and Neuroscience. See [Jain 1999, Schaeffer 2007, Buluc et al. 2014] for related reviews.

Graph Clustering and Astronomy

One of the most important and challenging areas of research in modern Astrophysics and Cosmology concerns the study of the sequence of events that led to the formation of galaxies and the supermassive black holes that usually reside at their centres. Understanding of the complex astrophysics which led to the observed distribution of galaxies in space both in terms of properties and numbers is usually sought within the framework of the standard ΛCDM cosmological model (e.g. [Lacey et al. 2016]).

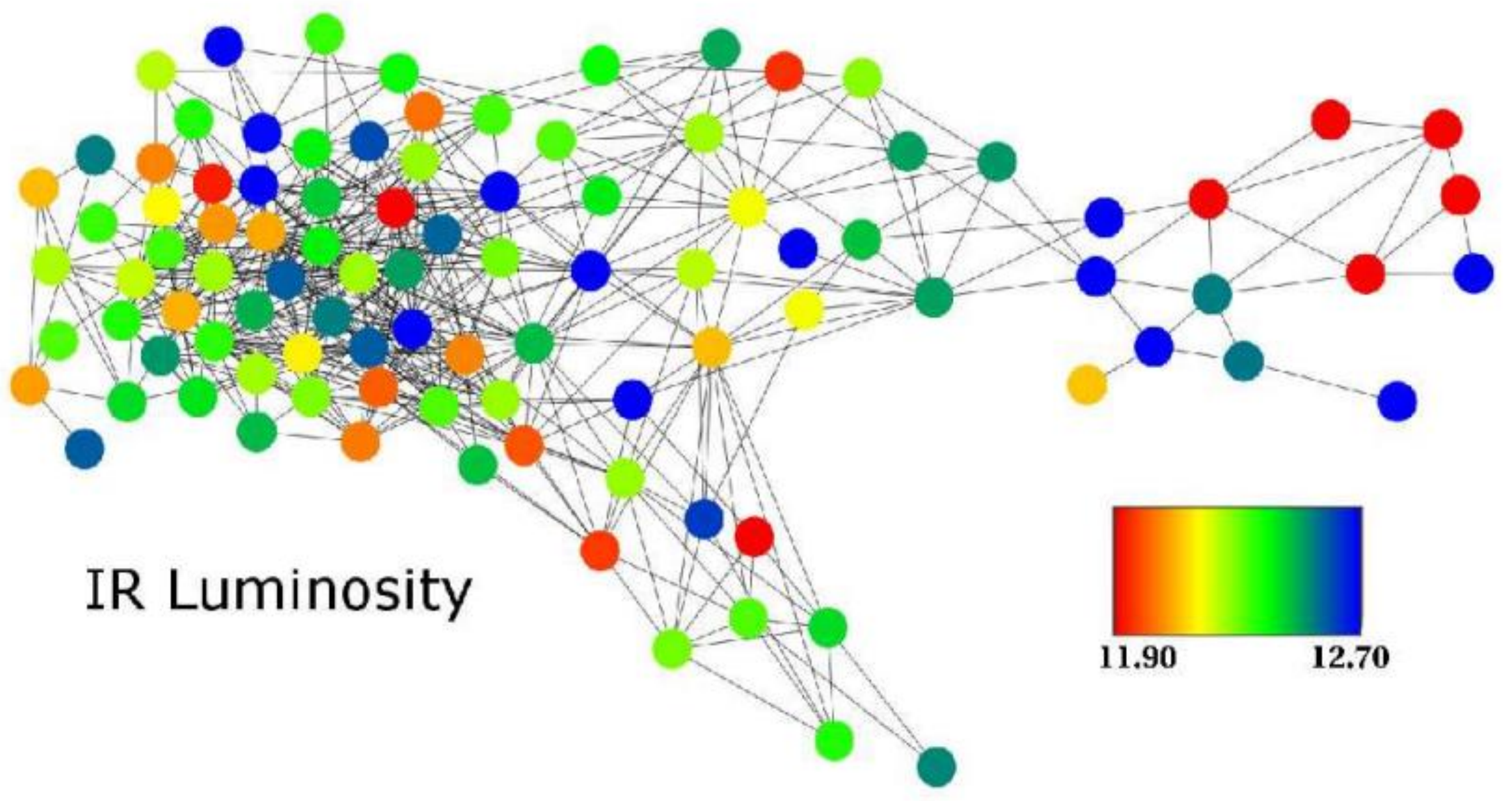
It is now clear that in order to understand the numerous processes that govern galaxy formation and evolution (quiescent star formation, bursts of star formation and accretion onto supermassive black holes) we need multi-wavelength or panchromatic observations of galaxies at all cosmic epochs.

Graph Theory has proven to be a powerful multidisciplinary tool to study networks and clustering in data. In our research project we make use of Graph Theory and Bayesian inferencing to study galaxy evolution, by applying them to the fields of Astrophysics (real ULIRG data) and Cosmology (galaxy formation simulations).

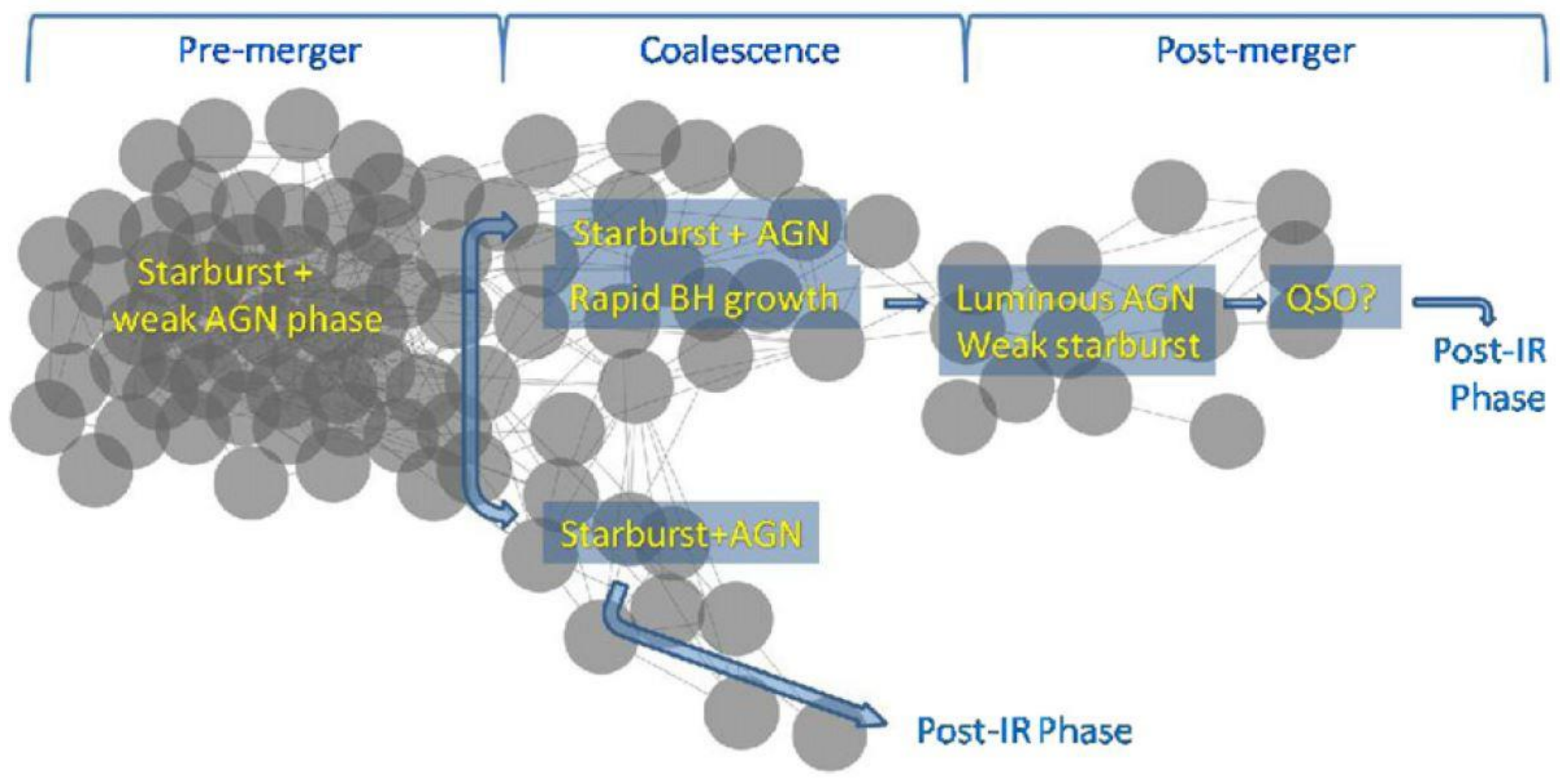
Clustering in Astrophysics

In the field of Astrophysics, we aim to identify and study the different stages of the evolution of Ultraluminous infrared galaxies so that an evolutionary paradigm can be developed, by using Graph Theory and Bayesian inferencing. The use of Graph Theory tools like *Cytoscape* and *neo4j* will enable us to group galaxies into different evolutionary stages based on their observed properties and in particular their mid-infrared spectra, such as black hole mass, characteristic PAH emission and IR luminosity, following the work of *Farrah et al. (2009)*, on low-redshift galaxies (z<0.4). Galaxies act as nodes and their characteristic similarities will act as a basis for the connections between them (edges), by using Bayesian inferencing. We will analyze infrared data taken by NASA's Spitzer Space Telescope in combination with other multi-wavelength data to determine certain properties of the evolutionary stage of these galaxies, such as AGN and starburst activity. Graph Theory could also be utilized to analyze future data obtained from JWST and SPICA.

Examples of Graph Theory applied to Astrophysics



IR Luminosity graph from a sample of 102 low-redshift galaxies
Farrah et al. (2009)



Schematic of the evolutionary paradigm discussed in Farrah et al. (2009)

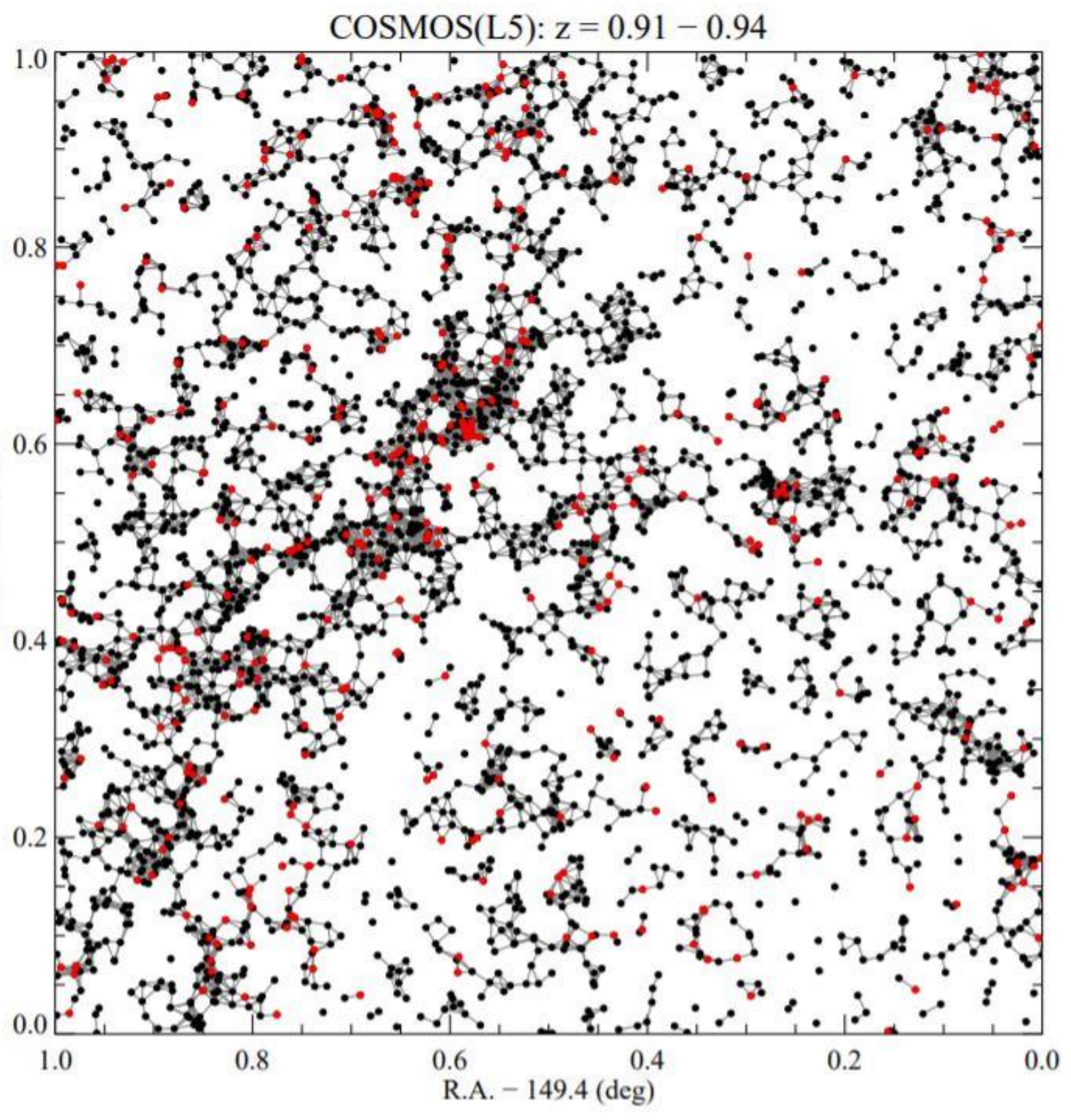
By performing graph theoretical clustering analysis we aim to study a large sample of galaxies at different redshifts, group them into different sets based on the observed characteristics displayed in different stages of their evolution and establish causal relations between the processes taking place. Our goal is to further constrain these key stages' features and ultimately create an evolutionary paradigm for ULIRGs. We aim to expand the scope of previous research done in the field, by utilizing data from higher redshifts and including more recent and better resolved galactic data. It is therefore imperative that we also make use of high performance computing to perform these tasks.

Clustering in Cosmology

In the domain of Cosmology, we will use Graph Theoretical tools to study the clustering of galaxies which as is well known contains important constraints to cosmological models and parameters (e.g. [White & Frenk 1991]). As shown by [Hong et al 2016] an analysis of the clustering of galaxies using graph theory can be much more informative about the topology of galaxy distribution compared to more traditional methods, like the two-point correlation function. We plan to use this approach for studying the clustering of galaxies in the state-of-the-art cosmological hydrodynamical EAGLE simulations, which are publicly available [McAlpine et al 2016], but also in observations of the real Universe by using data which are either already available (HerMES) or which will soon be made available with large multi-wavelength surveys of the sky (e.g. the FP7 funded Herschel Extragalactic Legacy Project or HELP in which members of this team participate see [Vaccari et al.2016], DES, LSST).

Related Work

Hong et al (2016) presented a novel method of studying the galaxy distribution which generates a graph or a network composed of vertices (nodes) and edges (links) from a galaxy distribution. They show that this method can distinguish between two galaxy distributions which have the same two-point correlations but different topologies. Hong et al (2016) used *igraph*, a collection of open source network analysis tools which can be programmed in R, Python and C (http://igraph.org/redirect.html) to study the clustering of simulated galaxies in the Illustris simulation (Vogelsberger et al. 2014).



References

[1] Efstathiou, A., Christopher, N., Verma, A., Siebenmorgen, R., 'Active galactic nucleus torus models and the puzzling infrared spectrum of IRAS F10214+4724', 2013, MNRAS, 436, 1873.
[2] Efstathiou, A., et al., 'Herschel observations and a model for IRAS 08572+3915: a candidate for the most luminous infrared galaxy in the local (z < 0.2) Universe', 2014, MNRAS, 437, L16.
[3] Santo Fortunato, 2010, Community detection in graphs, Physics Reports, Volume 486, Issues 3–5, February 2010, Pages 75-174.
[4] Farrah et al., 'An evolutionary paradigm for dusty Active Galaxies at low redshift', 2009, The Astrophysical Journal, 700:395–416.
[5] Bullmore and Sporns, 'Complex brain networks: Graph theoretical analysis of structural and functional systems', 2009, Nature Reviews Neuroscience, 10(3):186-98.
[6] David A. Bader et al., 'Graph Partitioning and Graph Clustering', 2015, Atlanta, GA, American Mathematical Society
[7] Lacey et al., 'A unified multiwavelength model of galaxy formation', 2016, MRNAS.462.3854L.
[8] White & Frenk 1991, 'Galaxy formation through hierarchical clustering', 1991, The AstrophysicalJournal, 379:52-79,1991 September 20.
[9] Hong et al., 'Discriminating Topology in Galaxy Distributions using Network Analysis', 2016, MNRAS, 459, 2690.
[10] McAlpince et al., 'The EAGLE simulations of galaxy formation: public release of halo and galaxy catalogues', 2016, Astronomy and Computing (2016), pp. 72-89.
[11] Vogelsberger et al., 'Properties of galaxies reproduced by a hydrodynamic simulation', 2014, Nature, volume509, pages177–182.