UNIVERSITY OF BIRMINGHAM

DOCTORAL THESIS PROPOSAL

Extracting Knowledge from Complex Networks

Author:
David MCDONALD

Supervisor:
Dr. Shan HE

A thesis proposal submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Department of Computer Science

August 9, 2017

"I'm fascinated by the idea that genetics is digital. A gene is a long sequence of coded letters, like computer information. Modern biology is becoming very much a branch of information technology."

Richard Dawkins

University of Birmingham

Abstract

Computer Science
Department of Computer Science

Doctor of Philosophy

Extracting Knowledge from Complex Networks

by David McDonald

Throughout nature, we observe complex systems everywhere – from the social interactions of individuals within society, right down to the interrelation and cooperation of proteins, genes and metabolites within the cell. When modelling these systems, sometimes it is convenient to represent them as networks of different entities and relationships. So the problem of understanding these systems is the problem of building sufficiently powerful network models and devising methodologies to extract useful features from them. This thesis proposal will focus upon these challenges within the context of computational biology – namely, how can we make best sense of the vast amounts of data produced in the post-genome age? We propose novel information extraction techniques from multilayer networks – representing the intercellular world – in the form of network embedding and module identification techniques. These models will provide an insight into the formation of these complex systems and help to guide bioscientists by offering real, testable hypotheses in the form of new interaction and functional predictions.

Contents

List of Figures

Chapter 1

Introduction

Most complex systems observed in nature can be represented as a complex network. A complex network is a highly heterogeneous object that is represented as a mathematical graph that is comprised of non-trivial and non-uniform features.

Despite being characterised as complex objects, we have observed that many real world networks, from protein interaction networks, to the internet, possess shared features that unite such seemingly disparate subject and give rise to the prevailing popularity of their study. For example, we have observed that these complex networks follow scale-free distribution of node degree (often with exponent $\gamma \in [2, 3]$) (Barabási and Albert, 1999; Barabási, 2009). Connections are preferentially made between nodes with a probability proportional to their existing degree, giving rise to the so-called preferential attachment model, a model that adheres to the old 'popularity is attractive' adage. Furthermore, complex networks also are characterised by the 'small world' phenomenon, where one would expect a small average shortest path length and a high degree of clustering (Watts and Strogatz, 1998); and are typically very sparse with the number of edges in the same order as the number of nodes.

The work proposed here will focussed primarily upon the study of complex networks within the context of computational biology. Given the explosion of data available to us thanks to modern experimentation techniques, the need for modelling such data, has never been greater. Through the medium of complex networks, researchers wish to simplify and represent the raw data in a data-driven and scientifically justified way, such that is becomes useful for the purposes of making functional predictions and tracing the causes of observed disease phenotypes.

Already, we have seen the application of complex network models to the problems of identification of biologically relevant modules according to genetic expression profiles (Ideker et al., 2002); the multi scale detection of the function and structure of brain networks (Ashourvan et al., 2017); and testable protein functional predictions (Palla et al., 2005; Zhang et al., 2016). And with the growing availability of data and drive of the network science community, models are becoming more and more complex, with the integration of multiple types of data and more sophisticated analysis techniques. It is the hope of the work proposed here to further this burgeoning field by proposing novel, integrative network models that allow for the extraction of testable hypotheses based on the data currently available to us.

Chapter 2

Literature Review

Network science is the interdisciplinary endeavour of making sense of the complex networks that we observe in nature. It unites scientists from such varied fields as mathematics, computer science, medicine, biology and sociology. Researchers from all these fields and more have contributed their expertise and perspective. This section aims to give a brief overview of some of the interesting areas of research happening in this field. For the purpose of brevity, the review presented here will only focus on the fundamental concepts that will prove to be relevant in the work proposed here. For a more general overview of network science, the reader is pointed towards Barabási, 2002, a book written by Albert-László Barabási, a well respected and prolific figure in this field.

2.1 Community Detection

Many real-world networks contain subsets of nodes that contain a higher degree of inter-connectivity than the rest of the network (Girvan and Newman, 2002; Palla et al., 2005; Lancichinetti, Fortunato, and Kertész, 2009). These subsets are commonly referred to as communities. A rigorous definition for a 'community' within a network still seems to elude the scientific community(Lancichinetti, Fortunato, and Kertész, 2009). However, the most popular definition among scholars is the planted l-partition model. This was popularised thanks to Girvan and Newman in their seminal work Girvan and Newman, 2002 and states that as long the probability of a node being connected to its group is greater than the probability of it being connected to the rest of the graph, then those groups are communities. 'Community detection' is the name given to the problem of finding the underlying community structure in a given network (Girvan and Newman, 2002). For example, groups of friends in a social network, functional modules in Protein-Protein Interaction (PPI) networks and scientific disciplines in co-authorship networks.

But, a general community detection algorithm does not yet exist. Many existing algorithms suffer from a number of issues. To name a few: The number and scale of communities must be known a-priori, which in most real applications, is infeasible. Additionally, the relationships between communities, both one the same level and at different ones, is lost. Identifying not only the community itself, but its position in the network as a whole, provides further insight into the often abstract interactions that comprise complex networks and so preserving this information when analysing a network is paramount. And, in some cases, the algorithms cannot deal with special cases: for example, modularity-based methods suffer from the so-called 'resolution limit' (Fortunato and Barthelemy, 2007).

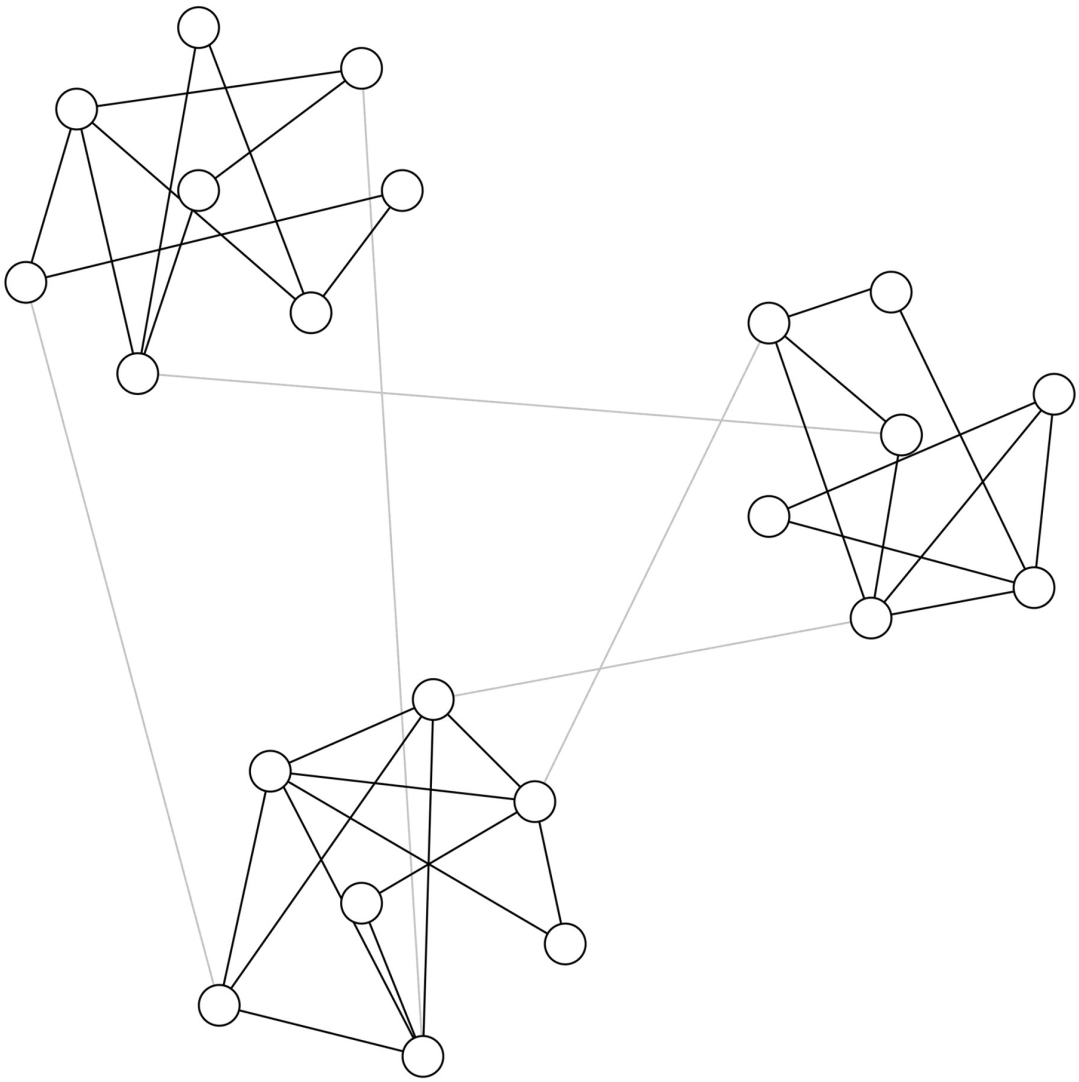


FIGURE 2.1: Representation of a network with three communities. The network has a higher proportion of edges between nodes of the same community than edges between nodes in different communities. Figure taken from Girvan and Newman, 2002.

2.1.1 Cut-Based and Spectral Approaches

The flagship Kernighan-Lin algorithm (Kernighan and Lin, 1970) focused on 'cutting' the network into modules, in such a way that the number of edges cut was minimized. However, this often favoured cuts of small, peripheral subgraphs, so it was adapted into ratio cut (Wei and Cheng, 1991), normalised cut (Shi and Malik, 2000) and min-max cut (Ding et al., 2001) that took the number of nodes in each resulting sub-graph into account, and thus resulted in a partition that was more balanced.

Contemporary cut-based approaches are concerned more with edges, rather than vertices and gave rise to a new measure for a good cut, called conductance. Conductance is still prolific in the literature: it has been used to detect communities in bipartite networks (Barber, 2007), combined with PageRank (Andersen, Chung, and Lang, 2006) and used as the basis for a greedy optimisation algorithm (Lancichinetti and Fortunato, 2009) capable of finding overlapping communities at different scales.

Spectral clustering dates back to the work of Donath and Hoffman in 1973 (Donath and Hoffman, 1973). However, it was popularized in the early 2000s (Shi and Malik, 2000; Ng, Jordan, Weiss, et al., 2002; Ding, 2004). Spectral methods rely upon constructing Laplacian matrices from the raw network data and eigen-decomposing them. Clustering the resulting eigenvectors results in clusters of the original data points. Spectral approaches have many advantages over other techniques and, as a result, they have become popular in the machine learning community for clustering on non-linear manifolds. According to Von Luxburg, 2007, 'these methods do not make assumptions about the form of the clusters' and are capable of correctly identifying typically challenging clusters, such as the famous two spirals example. For community detection, they have the additional benefit of efficiency, especially if the graph adjacency matrix is sparse.

2.1.2 Modularity-Based Approaches

The seminal work of Girvan and Newman (Girvan and Newman, 2002) marked a significant advance in the field by providing the first quantitative measure of a community: modularity. The modularity of a partition of a network defined as

$$Q = \sum_{s=1}^{m} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right] \tag{2.1}$$

scores a network partition by comparing the number of links inside a given module with the expected number that would be found in a random graph of the same size and degree sequence. Here, m is the number of modules in the partition, l_s is the number of links in module s, L is the total number of links in the network and d_s is the total degree of the nodes in s. Girvan and Newman propose a hierarchical divisive algorithm that removes edges based on their 'betweenness' (the number of shortest paths from two nodes in the network that go through them) until the modularity quality function is maximized. The early work of Girvan and Newman has since been expanded upon. For example, edge clustering in favour of edge-betweenness (Radicchi et al., 2004), iteratively adding links to a module based on their expected increase in modularity (Clauset, Newman, and Moore, 2004), and multi-stage local optimization in the popular Louvain algorithm (Blondel et al., 2008).

2.1.3 Flow-Based Approaches

The Markov Clustering algorithm (MCL) simulates a diffusion process on a graph by repeatedly performing stages of expansion and inflation and only keeping the k largest elements for efficiency (Van Dongen, 2001).

Another significant flow-based approach is the the work of Rosvall and Bergstrom with their famous Infomap algorithm (Rosvall and Bergstrom, 2007) that translated the problem of community detection into the problem of optimally compressing the information in a graph such that the most information can be uncovered when the compression is decoded. They used simulated annealing to minimize a function that represented both compression and data loss resulting in a map that "best captures the community structure with respect to the dynamics on the network" (De Domenico et al., 2015). While slow and computationally expensive, this approach was also shown to work well with dynamic processes in their later work (Rosvall and Bergstrom, 2008).

2.1.4 Deep Learning

The deep learning community has began to explore the possibilities of using neural networks for clustering in the graph domain. Convolutional neural networks (CNNs), powerful machine learning tools that have proven very successful for challenging classification tasks that have recently been generalised to take a graph input (Defferrard, Bresson, and Vandergheynst, 2016). CNNs have also been used for semi-supervised learning on graphs, where the they are capable of learning both graph structure and node features (Kipf and Welling, 2016).

2.1.5 Hierarchical Community Detection

In many networks representing complex real-world phenomena, finding a single cover – where each node is assigned to exactly one community – does not accurately reflect the underlying community structure of the data being represented. Sometimes, nodes can belong to more than one community, and sometimes communities overlap. The first algorithm to consider overlapping communities was CFinder in 2006 (Adamcsek et al., 2006). Drawing from the earlier work of Palla et al. and the Clique Percolation Method (CPM) (Palla et al., 2005), CFinder considered communities as the unions of k-cliques and so rolled k-cliques across the graph to detect communities. While computationally expensive, it was able to deal with overlapping cases, and opened the door for further study. Shen et al. proposed EAGLE in 2009 (Shen et al., 2009) that used maximal cliques, an agglomerative hierarchical structure and a modified modularity quality function that detected complex overlapping community structures.

The hierarchical nature of modularity-based clustering methods can allow them to detect communities at different scales. Lancichinetti, Fortunato, and Kertész, 2009 used local optimization to maximize a fitness function with a parameter that controlled the size of communities detected. Other work includes multi-scale quality functions that can uncover hierarchical communities and produce several different partitions of a graph, the post-processing of clusters found by hierarchical methods (encoded in a dendogram) (Pons and Latapy, 2011), and Bayesian non-negative matrix factorisation that performs 'soft-partitioning' and assigns node participation scores to modules (Psorakis et al., 2011).

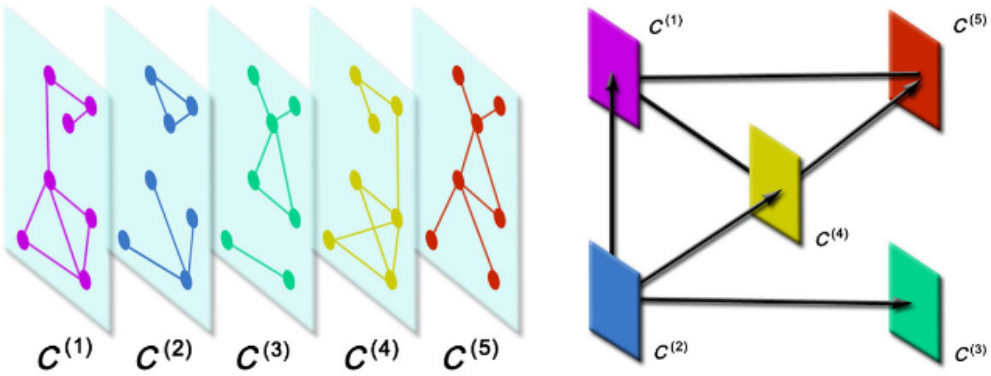


FIGURE 2.2: A schematic representation of a multilayer complex network. Here, $C^{(n)}$ is the nth layer of the network. The right hand figure shows how the layers connect to each other, forming a network of networks. This network would be represented as a four-dimensional tensor with each entry $a_{ij}^{\alpha\beta}$ corresponding to the strength of link form node i in layer α to node j in layer β. Figure taken from De Domenico et al., 2013.

2.1.6 Multilayer Community Detection

When more than one type of actor takes part in a complex system, a single layer network cannot accurately capture this. In these cases, one may construct a model consisting of several connected 'layers' of networks, one for each type of actor in the system (Kivelä et al., 2014). These models allow for nodes to connect to nodes in the same layer but also to nodes in other layers through special relationships. For example, one may construct a multilayer network consisting of layers of genes, connected by co-expression, proteins connected by physical interaction and inter-layer links between genes and proteins if a protein is known to related that genes expression levels. When one introduces enough types of layers, and inter-layer connections, one can consider a multilayer network to be a sort of 'network' of networks.

Community detection within these models follows from the same principles as the single layer case. In fact, many classical community detection algorithms have been generalized for application to multiplayer networks (Mucha et al., 2010). For example, Ashourvan et al., 2017 used a multi scale variation of the Louvain algorithm to detect hierarchical communities in functional brain networks, and De Domenico et al., 2015 generalizes random walking and the map equation (Rosvall and Bergstrom, 2008) to move across multiple layers.

These algorithms can successfully identify closely related sets of nodes, of different types, which, in the context of computational biology can be useful to predict function and missing links. Special cases of multilayer networks are the multi-slice networks where the nodes are the same for each layer but the topology of the network may differ. Examples include, temporal networks where each layer represents a different time point and networks representing many different types of interactions amongst nodes.

2.2 Active Module Identification

While community detection only considers only the structure of the network at hand, often nodes in network may be enriched with additional attributes that are

not solely based upon the observed topology of the network. For example, people in a social network may be annotated with preferences such as hobbies and interests and we would expect two people with the same interests to still somehow be similar, even if we do not observe a direct link between them in the network. Within the context of computational biology, this is perhaps even more relevant, due to the vast quantity and variety of data now available to us, and the successes of integrative models in the past. Integrative models get their name from the principle of integrating observed data (say, gene expression) with prior knowledge (often in the form of a known protein interaction network and/or previously curated functional annotations). Mitra et al., 2013 offers a summary of many the integrative approaches popular in the literature.

One of the most successful integrative approach is the identification of so-called 'active modules'. It is a relatively recent trend within the interdisciplinary fields of network science and translational medicine and aims to augment known physical interactions with observed expression levels to identify connected sub-graphs (called sub-networks for the remainder of this proposal) that are maximally differently expressed. Ideker et al. was the first to formalise this problem in 2002 (Ideker et al., 2002). Given a known PPI network G and a matrix of gene expression levels with their corresponding p-values P, we compute a z-score for each gene i in the network as:

$$z_i = \Phi^{-1}(1 - p_i) \tag{2.2}$$

and then score identified sub-networks A in an aggregated manner with

$$z_A = \frac{1}{\sqrt{k}} \sum_{i \in A} z_i \tag{2.3}$$

where a high z_A represents a biologically active sub-network. Here, Φ^{-1} is the inverse normal CDF. While the problem is related to community detection – the sub graphs must be connected – but biologically relevant modules often do not align with communities based solely on network topology.

Computing an exact solution is NP-hard (Ideker et al., 2002), so the authors employ a heuristic search based on simulated annealing to search for the maximally scoring sub-graph in the network. Genetic algorithms (GAs) (Klammer et al., 2010), greedy methods (Nacu et al., 2007) and propagation of flow from cancer genes have since been used (Vandin, Upfal, and Raphael, 2011). More recent work has employed a memetic algorithm to ensure connectedness (Li et al., 2017a); a multi-objective optimisation process to control the trade off between biological activity and functional enrichment of the detected modules (Chen, Liu, and He, 2017); and a cooperative co-evolutionary approach (He et al., 2016). Interestingly, despite the NP-hardness of the problem, Dittrich et al., 2008 showed that by transforming the above problem into the well known Prize Collecting Stein Tree (PCST) problem, exact solutions can be obtained in reasonable computational time with integer programming.

2.3 Network Embedding

Several models in the literature assume the existence of an underlying metric space that controls the topology of the network. They suppose that entities that are closer together in this space are more 'similar' and have a higher probability of being connected. These models aim to infer the geometry of these spaces and the positions

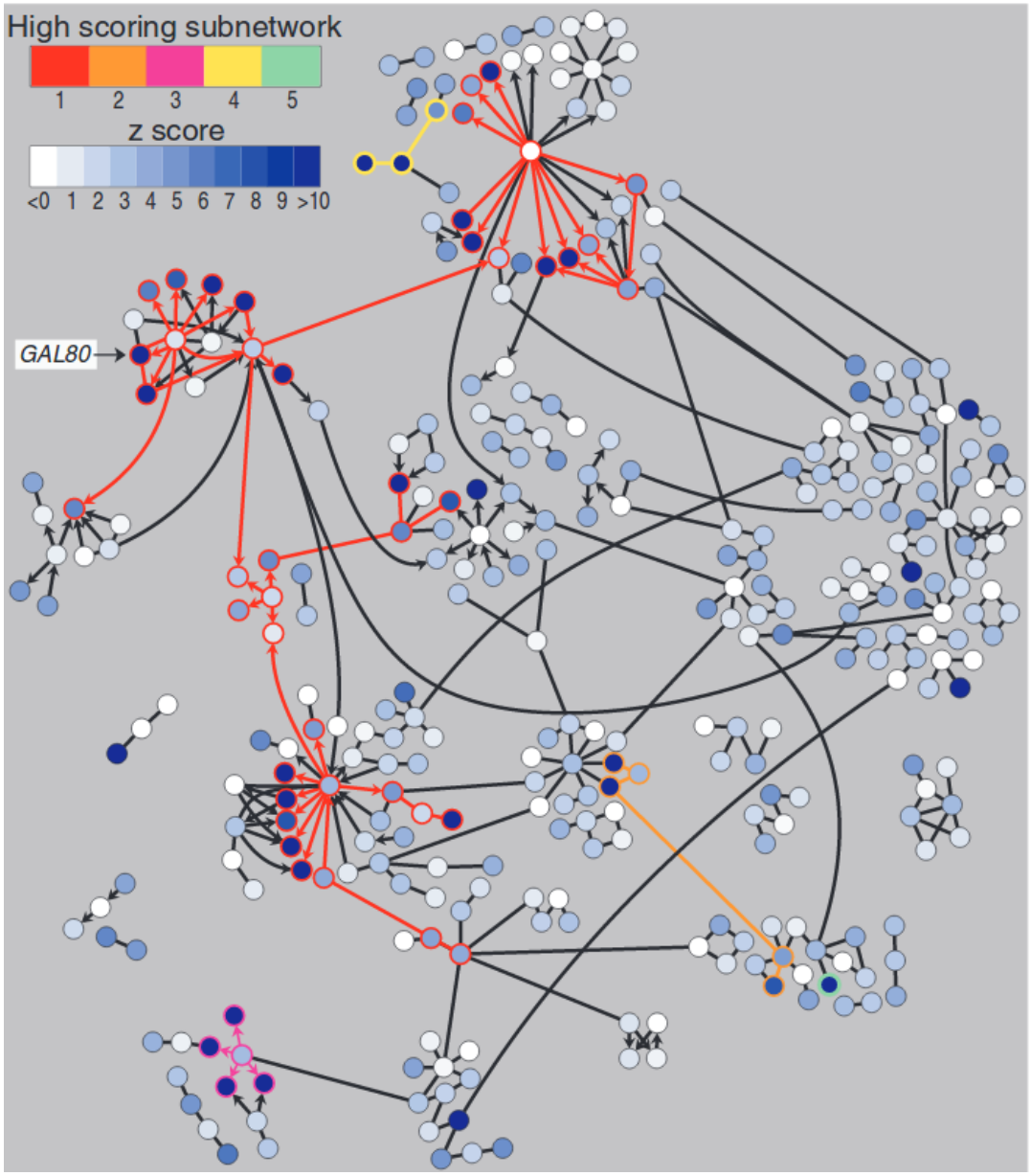


FIGURE 2.3: Top five active modules identified by a simulated annealing approach in Ideker et al., 2002. Nodes in this network represent genes, undirected edges signify that the proteins encoded by these genes physically interact and a directed edge represents transcription influence. Nodes are colours by the likelihood of their observed expression change in a GAL80 knockout experiment. The subnetworks identified represent the most 'biologically active' modules in the network. Figure taken from Ideker et al., 2002.

of nodes within the space, such that the probability of reconstructing the observed network is maximised, for purposed of better understanding of the system and visualisation. This is the so-called network embedding, and is the cornerstone of the field of *network geometry*.

Network embedding is closely related to the field of manifold learning. Indeed, many classical non-linear manifold learning techniques, such as Isomap (Tenenbaum, De Silva, and Langford, 2000) and Laplacian Eigenmaps (Belkin and Niyogi, 2002), must first construct nearest neighbour graphs based on dissimilarities between samples before dimensionality reduction takes place. Many of these techniques are directly applicable to embedding of complex networks by simply omitting the graph construction step.

An interesting and popular embedding paradigm in the literature comes from natural language processing. In particular, the Skipgram model and the Word2Vec algorithm that aims to vectorise words and phrases in a semantic space such that similar words are mapped close together (Mikolov et al., 2013a; Mikolov et al., 2013b). The principle idea is, given a corpus of words and a particular sentence, generate a 'context' for each input word with the aim of maximising the likelihood of observing context words in the embedding space, given the input word. Similarities are measures by dot products and accordingly, observation probabilities are computed using a multilayer perception with a linear hidden layer and softmax output. Through the use of sub-sampling and negative sampling (replacing softmax with sigmoid), training can be made very efficient and the resulting embeddings can be obtained from the activation of the hidden units. This idea naturally extends to networks, where sentences are replaced by 'neighbourhood graphs' generated from random walks. Furthermore, the shallow architecture of the Skipgram model has been replaced with multiple non-linear layers to learn the highly non-linear relationships between nodes (Perozzi, Al-Rfou, and Skiena, 2014; Tang et al., 2015). By introducing additional parameters into the random walk to control a breadth vs. depth first neighbourhood search, Grover and Leskovec, 2016 were able to identify neighbourhoods of nodes with high *homophily* and high structural similarity.

2.3.1 Embedding to a Hyperbolic Metric Space

An emerging popular belief in the literature is that the underlying metric space of most complex networks is in fact hyperbolic. Nodes in real world networks often form a *taxonomy*, where nodes are grouped hierarchically into groups in an approximate tree structure. Hyperbolic spaces can be viewed as continuous representations of this tree structure and so models that embed networks into hyperbolic space have proven to be increasingly popular in the literature (Krioukov et al., 2009; Krioukov et al., 2010). In fact, this assumption has already had proven success in the task of greedy forwarding of information packets where nodes use only the hyperbolic coordinates of their neighbours to ensure packets reach their intended destination (Papadopoulos et al., 2010).

The most popular of all these models is the Popularity-Similarity (or PS) model (Papadopoulos et al., 2011). This model extends the "popularity is attractive" aphorism of preferential attachment to include node similarity as a further dimension of attachment. Nodes like to connect to popular nodes but also similar ones. The PS model sustains that the clustering and hierarchy observed in real world networks is the result of this principle (Alanis-Lobato, Mier, and Andrade-Navarro, 2016a). This model has a simple interpretation in two dimensional hyperbolic space, where

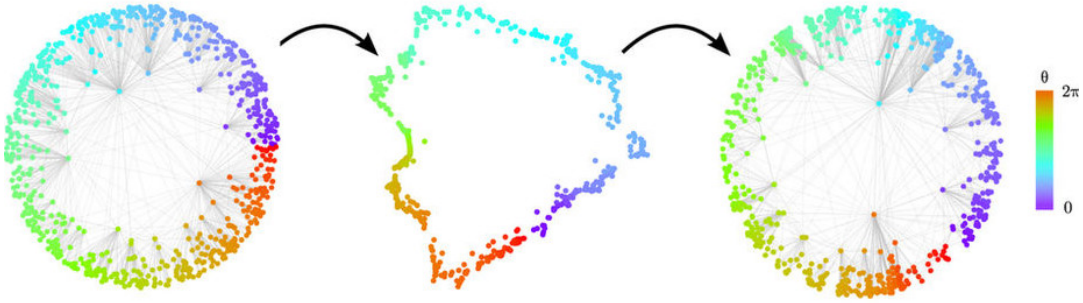


FIGURE 2.4: An artificial complex network embedded onto the two dimensional hyperbolic disk. (a) an artificial network generated according to the PS model. (b) the LABNE algorithm in progress to determine each nodes radial coordinates according to the graph Laplacian. (c) determining radial coordinated by popularity (node degree). Figure taken from Alanis-Lobato, Mier, and Andrade-Navarro, 2016a.

nodes are placed on a hyperbolic disk, with radial coordinates representing popularity and angular coordinates representing similarity. Then the hyperbolic distance between two nodes $\mathbf{x}_1 = (r_1, \theta_1)$ and $\mathbf{x}_2 = (r_2, \theta_2)$, given by the hyperbolic law of cosines:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \text{arccosh}(\cosh(r1)\cosh(r2) - \sinh(r1)\sinh(r2)\cos(\Delta\theta)) \tag{2.4}$$

$$\Delta\theta = \pi - |\pi - |\theta_1 - \theta_2|| \tag{2.5}$$

controls their connection probabilities. Nodes with short hyperbolic distances show a higher probability of being connected.

Maximum likelihood (ML) was used in Papadopoulos et al., 2011 to search the space of all PS models with similar structural properties as the observed network, to find the one that fit it best. This was extended by the authors in Papadopoulos, Psomas, and Krioukov, 2015; Papadopoulos, Aldecoa, and Krioukov, 2015. Due to the computationally demanding task of maximum likelihood estimation, often heuristic methods are used. For example, Alanis-Lobato, Mier, and Andrade-Navarro, 2016a used Laplacian Eigenmaps to efficiently estimate the angular coordinates of nodes in the PS model. The authors then combined both approaches to leverage the performance of ML estimation against the efficiency of heuristic search with a user controlled parameter in Alanis-Lobato, Mier, and Andrade-Navarro, 2016b. Additionally, Thomas et al., 2016 propose the use of classical manifold learning techniques in the PS model setting with a framework that they call *coalescent embedding*.

2.3.2 Embedding with Node Attributes and Dynamics

Little work has been done to embed networks, accounting for both topology and node attributes, however recent years has shown it to become more prolific. Gibert, Valveny, and Bunke, 2012 embed into a vector space based on the statistics of attributes and pairs of attributes, Li et al., 2017b draw from the well known fields of manifold learning and multi-view learning to align the projections based on topology and attributes and Liao et al., 2017 use deep learning. In Niepert, Ahmed, and Kutzkov, 2016, the authors generalised convolutional neural networks from regular pixel lattices to arbitrary graphs. It is worth noting that by transforming an unweighted graph into a so-called 'flow graph' (Lambiotte et al., 2011) by weighting links by node expression, many embedding techniques that are applicable to

weighted graphs can be applied to unweighted graphs with node attributes. However, it is not clear how to do this if the graphs are already weighted, or nodes are annotated with discrete or multiple attributes.

Chapter 3

Proposed Work

3.1 Research Questions

After a thorough literature review, the following research questions have been identified as the focus of the upcoming PhD:

1. Can a multi-layer network model be constructed such that the hyperbolic metric space that underpins complex network formation be uncovered for multi-layer networks?
2. What interpretable information, such as disease modules or functional predictions, can be extracted from such a model?
3. My constructing data-driven models, can we predict protein abundance and phosphorylation?

3.2 Network Model Construction

When constructing multilayer network models, the main challenge is the weighting of inter-layer links. Intra-layer links (such as protein-protein or metabolite-reaction links) are simple and constructed directly from the data. For example, we will place a link between proteins that physically interact or a link between a metabolite and a reaction if that metabolite is involved in that reaction, either as a *substrate* or *product*. Inter-layer links are placed between entities if they are related – for example, a protein is part of a metabolite or is associated with a particular phenotype.

3.3 Network Embedding to Hyperbolic Metric Space

Network embedding can be considered a preprocessing step for knowledge extraction from biological data. Knowledge such as disease module identification, for example. Embedding allows for the transformation of problems: from module to detection to clustering. But, selecting a metric space and embedding method is a very non-trivial problem.

The recent trend towards models that support an underlying hyperbolic metric space is very appealing, as results are extremely positive, and the space is a generalisation of the hierarchical structure that we know actually exists in complex networks. However, there are a number of questions as yet unanswered that this work shall focus on. Firstly, how to embed networks with node attributes into hyperbolic space, such that active modules can be identifies as clusters. While single continuous node attributes can be combined with into the topology of an unweighted graph

via construction of flow graphs (Lambiotte et al., 2011), some of the current hyperbolic embedding methods (for examples LABNE Alanis-Lobato, Mier, and Andrade-Navarro, 2016a) is unsuitable for weighted graphs. Furthermore, flow graphs cannot be used if there is more than one node attribute. Another challenge is embedding of multilayer networks.

This work will first attempt to tackle both of these problems for hyperbolic network embedding with the application of multi-view learning techniques. We will consider each layer as a separate view and align the separate embeddings onto the same space using, for example, kernel canonical correlation analysis (KCCA) (Hardoon, Szedmak, and Shawe-Taylor, 2004). In the case of networks with node attributes, we shall first construct an 'attribute network' by connecting nodes with similar attributes, for example with KNN. Another technique that shall be adapted for the hyperbolic space is the recently proposed extension of convolutional neural networks to arbitrary graphs (Niepert, Ahmed, and Kutzkov, 2016). Additional ideas for future work include extending the currently used 2D hyperbolic plane to three dimensions to allow for simultaneous embedding of all layers. However, very little thought has been put into this currently.

3.4 Module Identification and Knowledge Extraction

The identification of modules has proven to be an important step in extracting knowledge from a complex network. It is clear from the literature that topological modules (communities) rarely align with complete functional modules, such as pathways (He et al., 2016). Because of this, the addition of prior knowledge is essential to guide the module search to look beyond topology alone. Similarly, the identification of active modules must be guided by both node attributes (gene expressions) and network topology (identified sub-networks must be connected) (Ideker et al., 2002). Furthermore, modules in biological networks have been shown to have a highly hierarchical structure – where large modules of more general function are composed of many smaller ones that perform more specific tasks. As such, any model for meaningful biological module recovery must take hierarchy into account. Also, biological systems are dynamic – in that their activity, and maybe even topology changes over time – and composed of many types of entities and interactions. Because of this, network models must be scalable from single to multiple layers in tractable time. Finally, due to experimental error and measurement difficulties often data is incomplete and so any model must be robust enough to handle this. All of these factors combined make the problem of identifying the modular structure of biological systems a complex optimisation and combinatorial problem. This work intends to overcome some of the weaknesses of other techniques – such as poor scalability or user-defined parameters to control the number of modules or module scale – by developing models that allow for the extraction of modules in a more data-driven way.

The first such model will use the topology preserving property of the self-organising map to construct a map of communities based on an observed network. The model will use the Growing Hierarchical Self-Organising Map (GHSOM) (Dittenbach, Merkl, and Rauber, 2000) to grow maps organically with no need to specify map size a priori. Furthermore, neurons in the map may be selected for expansion, resulting in new maps of higher granularity for those data points. This results in a set of maps with hierarchical structure that reflects the underlying hierarchical community structure of the data.

This model has a number of drawbacks, in that it is solely topology based, and requires some network embedding to a Euclidean (or perhaps hyperbolic) space or adaption of the algorithm to work directly on graphs, as in Yamakawa, Horio, and Hoshino, 2006. The map structure may also impose too strong an emphasis on module connection and so the Growing Hierarchical Neural Gas (Palomo and López-Rubio, 2016) may be a more suitable choice, as it only connections between neurons are allowed also form in a data-driven way. It is also unclear how to scale this module to multiple layers without utilising some multi-view learning techniques to align the embeddings of the layers into the chosen space. Also, despite the appealing nature of not needing to know the number of modules or structure a priori, GHSOM is controlled by a number of parameters that can be difficult to tune.

Future work will attempt to generalise heuristic algorithms for active module identification to networks with many layers, such that the identified modules contain nodes from multiple layers. In the context of dynamical data, this will identify modules that are conserved across time points. Existing algorithms for multi-scale community detection (Mucha et al., 2010; Ashourvan et al., 2017) could be combined with heuristics to guide a search to include node scores as well as topology.

3.5 Protein Abundance and Phosphorylation Level Prediction

Protein function almost always determines the biological function of a cell. As a result, proteomics (the study of proteins) is perhaps the most well studied area of computational biology. However, there are a number of challenges that face the proteomics research community today. Firstly, protein datasets are often incomplete, due to experimental errors and other challenges. As such, an method to effectively impute missing values would enable the community to access data that would otherwise be ignored in analyses. Furthermore, protein data is expensive to generate and not as readily available to the scientific community as Copy-Number Alterations (CNA) and mRNA data. It is well known that mRNA is used to transfer information to proteins, however it has been shown that RNA expression alone is weakly predictive of protein levels (Vogel and Marcotte, 2012). We would like to augment this analyses with the integration of CNA profiles – which have been shown to affect protein abundances (Zhang et al., 2016) – for more accurate protein abundance prediction. Finally, by further integration of known proteomics profiles, we would like to predict the phosphorylation levels of proteins, as this has been shown to be one of the key so-called Post-translational Modifications (PTMs) that affect protein, and therefore cell, function. These challenges form the basis for the NCI-CPTAC DREAM Proteogenomics Challenge that this author, along with a distinguished team of researchers, shall participate in.

Tackling the imputation challenge shall begin by performing a servery and comparison against ground truth of many common imputation techniques found in the literature. For example, sample means, matrix factorization, spectral regularization (Mazumder, Hastie, and Tibshirani, 2010), Multiple Imputation by Chained Equations (MICE), and an autoencoder with an adapted objective as in Beaulieu-Jones et al., 2016. The best algorithm shall be adapted to give improved performance.

Prediction of protein abundance and phosphorylation levels shall be tacked in a three stage way. First, by the application of simple machine learning techniques, such as regression models. Then, by constructing network models. Finally, by augmenting our models with the inclusion of domain knowledge than can be extracted

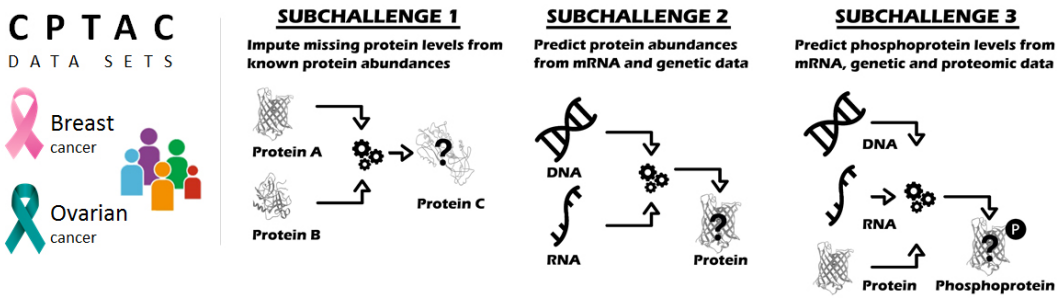


FIGURE 3.1: Schematic representation of the three sub-challenges ion the NCI-CPTAC DREAM Proteogenomics Challenge. Sub-challenge 1: impute protein levels. Sub-challenge 2: predict protein abundances. Sub-challenge 3: predict phosphoprotein levels.

from the given data, such as ribosome number and binding site information, provided by an expert in the field.

Chapter 4

Work Plan

Table **??** below details a detailed plan of the next year of work. The plan becomes much more general after than, as following work depends greatly upon the success of my early work and any new developments in the field. A Gantt chart generated from this table is available in the appendix.

Task Name	Start Date	End Date
Literature Review	07/28/16	08/16/17
Community Detection	07/28/16	01/18/17
Active Modules	01/19/17	05/03/17
Network Embedding	05/04/17	08/16/17
GHSOM	11/15/16	08/07/17
Reading	11/15/16	11/28/16
Construct Model	11/29/16	04/17/17
Results	04/18/17	05/29/17
Writing Paper	05/30/17	08/07/17
DREAM Challenge	08/01/17	11/14/17
Team Forming	08/01/17	09/11/17
Challenge Starts	09/12/17	09/12/17
Simple Machine Learning Techniques	09/13/17	10/03/17
Network Models	10/04/17	10/24/17
Integration of Domain Knowledge	10/25/17	11/14/17
Multilayer Hyperbolic Embedding Model	08/01/17	01/01/18
Prototyping and Collaboration	08/01/17	09/11/17
Model construction	09/12/17	11/20/17
Write Paper?	12/19/17	01/01/18
Return to GHSOM and Topological Module Detection	01/01/18	07/27/18
Multilayer Hyperbolic Embedding with Node Attributes/Dynamics	07/30/18	02/22/19
Module Identification	02/25/19	07/12/19
Write Thesis	05/01/19	02/04/20

TABLE 4.1: Timeline of proposed PhD. Gantt chart can be found in the appendix.

Chapter 5

Implications of Research

The main purpose of the work proposed here is to provide that can better make sense of all the data that is produced by bioscientists every day. It will generate knowledge from data through the means of extracting useful and interpretable features from complex network models that allow for the automatic generation of predictions or further study by domain experts. This will be achieved through novel network construction, embedding and feature extraction methods.

Hyperbolic network embedding has been shown to help uncover and understand how complex networks are formed (Krioukov et al., 2010). Furthermore, it has successfully predicted missing links in networks (Alanis-Lobato, Mier, and Andrade-Navarro, 2016a), which is particularly relevant to biosciences as testing for physical protein interactions is an expensive procedure that benefits from predictive models guiding the experimentation. Generalising these models to multiple levels may potentially allow for link prediction across different entities. For example, proteins and Gene Ontology terms that they have not yet been assigned to. Furthermore, multi-level hyperbolic embedding may be able to uncover a hierarchical structure that considers all entities in the system, rather than the structure of the entities amongst amongst themselves.

This will further be enhanced by more sophisticated module detection schemes. Schemes that are able to successfully harness to the topology of the network to detect subnetworks that not only show high biological activity but also high functional cohesion, while preserving the known hierarchical organisation of biological modules, in a data-driven way with few parameters to tune. This will allow for a deeper and more comprehensive understanding of how biological systems form and what role each member of the system plays.

We know that protein abundance is often indicative of cellular function, and yet proteomic profiling data from mass spectrometry based experiments often contain a large number of missing values due to the dynamic nature of the mass spectrometry instruments. This data is valuable and we would not wish to discard it out of hand. RNA/DNA is much cheaper to generate and more widely available than proteomic data, but RNA alone is only weakly predictive (Vogel and Marcotte, 2012). Development of a novel imputation method, based on machine learning and network science techniques would allow bioscientists to harness data to its full capacity with high confidence. And new protein abundance and phosphorylation level prediction techniques will result in improved biomarker development and avoid expensive enrichment testing.

Bibliography

Adamcsek, Balázs et al. (2006). "CFinder: locating cliques and overlapping modules in biological networks". In: *Bioinformatics* 22.8, pp. 1021–1023.

Alanis-Lobato, Gregorio, Pablo Mier, and Miguel A Andrade-Navarro (2016a). "Efficient embedding of complex networks to hyperbolic space via their Laplacian". In: *Scientific Reports* 6.

— (2016b). "Manifold learning and maximum likelihood estimation for hyperbolic network embedding". In: *Applied Network Science* 1.1, p. 10.

Andersen, Reid, Fan Chung, and Kevin Lang (2006). "Local graph partitioning using pagerank vectors". In: *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. IEEE, pp. 475–486.

Ashourvan, Arian et al. (2017). "Multi-scale detection of hierarchical community architecture in structural and functional brain networks". In: *arXiv preprint arXiv:1704.05826*.

Barabási, Albert-László (2002). "Linked: How everything is connected to everything else and what it means". In: *Plume Editors*.

— (2009). "Scale-free networks: a decade and beyond". In: *science* 325.5939, pp. 412–413.

Barabási, Albert-László and Réka Albert (1999). "Emergence of scaling in random networks". In: *science* 286.5439, pp. 509–512.

Barber, Michael J (2007). "Modularity and community detection in bipartite networks". In: *Physical Review E* 76.6, p. 066102.

Beaulieu-Jones, Brett K et al. (2016). "Missing data imputation in the electronic health record using deeply learned autoencoders". In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. Vol. 22. NIH Public Access, p. 207.

Belkin, Mikhail and Partha Niyogi (2002). "Laplacian eigenmaps and spectral techniques for embedding and clustering". In: *Advances in neural information processing systems*, pp. 585–591.

Blondel, Vincent D et al. (2008). "Fast unfolding of communities in large networks". In: *Journal of statistical mechanics: theory and experiment* 2008.10, P10008.

Chen, Weiqi, Jing Liu, and Shan He (2017). "Prior knowledge guided active modules identification: an integrated multi-objective approach". In: *BMC systems biology* 11.2, p. 8.

Clauset, Aaron, Mark EJ Newman, and Cristopher Moore (2004). "Finding community structure in very large networks". In: *Physical review E* 70.6, p. 066111.

De Domenico, Manlio et al. (2013). "Mathematical formulation of multilayer networks". In: *Physical Review X* 3.4, p. 041022.

De Domenico, Manlio et al. (2015). "Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems". In: *Physical Review X* 5.1, p. 011027.

Defferrard, Michaël, Xavier Bresson, and Pierre Vandergheynst (2016). "Convolutional neural networks on graphs with fast localized spectral filtering". In: *arXiv preprint arXiv:1606.09375*.

Ding, Chris (2004). "A tutorial on spectral clustering". In: *Talk presented at ICML.*

Ding, Chris HQ et al. (2001). "A min-max cut algorithm for graph partitioning and data clustering". In: *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. IEEE, pp. 107–114.

Dittenbach, Michael, Dieter Merkl, and Andreas Rauber (2000). "The Growing Hierarchical Self-Organizing Map." In: *IJCNN (6)*, pp. 15–19.

Dittrich, Marcus T et al. (2008). "Identifying functional modules in protein–protein interaction networks: an integrated exact approach". In: *Bioinformatics* 24.13, pp. i223–i231.

Donath, William E and Alan J Hoffman (1973). "Lower bounds for the partitioning of graphs". In: *IBM Journal of Research and Development* 17.5, pp. 420–425.

Fortunato, Santo and Marc Barthelemy (2007). "Resolution limit in community detection". In: *Proceedings of the National Academy of Sciences* 104.1, pp. 36–41.

Gibert, Jaume, Ernest Valveny, and Horst Bunke (2012). "Graph embedding in vector spaces by node attribute statistics". In: *Pattern Recognition* 45.9, pp. 3072–3083.

Girvan, Michelle and Mark EJ Newman (2002). "Community structure in social and biological networks". In: *Proceedings of the national academy of sciences* 99.12, pp. 7821–7826.

Grover, Aditya and Jure Leskovec (2016). "node2vec: Scalable feature learning for networks". In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 855–864.

Hardoon, David R, Sandor Szedmak, and John Shawe-Taylor (2004). "Canonical correlation analysis: An overview with application to learning methods". In: *Neural computation* 16.12, pp. 2639–2664.

He, Shan et al. (2016). "Cooperative co-evolutionary module identification with application to cancer disease module discovery". In: *IEEE Transactions on Evolutionary Computation* 20.6, pp. 874–891.

Ideker, Trey et al. (2002). "Discovering regulatory and signalling circuits in molecular interaction networks". In: *Bioinformatics* 18.suppl_1, S233–S240.

Kernighan, Brian W and Shen Lin (1970). "An efficient heuristic procedure for partitioning graphs". In: *Bell system technical journal* 49.2, pp. 291–307.

Kipf, Thomas N and Max Welling (2016). "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907*.

Kivelä, Mikko et al. (2014). "Multilayer networks". In: *Journal of complex networks* 2.3, pp. 203–271.

Klammer, Martin et al. (2010). "Identifying differentially regulated subnetworks from phosphoproteomic data". In: *BMC Bioinformatics* 11.1, p. 351.

Krioukov, Dmitri et al. (2009). "Curvature and temperature of complex networks". In: *Physical Review E* 80.3, p. 035101.

Krioukov, Dmitri et al. (2010). "Hyperbolic geometry of complex networks". In: *Physical Review E* 82.3, p. 036106.

Lambiotte, R. et al. (2011). "Flow graphs: Interweaving dynamics and structure". In: *PHYSICAL REVIEW E* 84.

Lancichinetti, Andrea and Santo Fortunato (2009). "Community detection algorithms: a comparative analysis". In: *Physical review E* 80.5, p. 056117.

Lancichinetti, Andrea, Santo Fortunato, and János Kertész (2009). "Detecting the overlapping and hierarchical community structure in complex networks". In: *New Journal of Physics* 11.3, p. 033015.

Li, Dong et al. (2017a). "Active module identification in intracellular networks using a memetic algorithm with a new binary decoding scheme". In: *BMC genomics* 18.2, p. 209.

Li, Jundong et al. (2017b). "Attributed Network Embedding for Learning in a Dynamic Environment". In: *arXiv preprint arXiv:1706.01860*.

Liao, Lizi et al. (2017). "Attributed Social Network Embedding". In: *arXiv preprint arXiv:1705.04969*.

Mazumder, Rahul, Trevor Hastie, and Robert Tibshirani (2010). "Spectral regularization algorithms for learning large incomplete matrices". In: *Journal of machine learning research* 11.Aug, pp. 2287–2322.

Mikolov, Tomas et al. (2013a). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*, pp. 3111–3119.

Mikolov, Tomas et al. (2013b). "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781*.

Mitra, Koyel et al. (2013). "Integrative approaches for finding modular structure in biological networks". In: *Nature reviews. Genetics* 14.10, p. 719.

Mucha, Peter J et al. (2010). "Community structure in time-dependent, multiscale, and multiplex networks". In: *science* 328.5980, pp. 876–878.

Nacu, Şerban et al. (2007). "Gene expression network analysis and applications to immunology". In: *Bioinformatics* 23.7, pp. 850–858.

Ng, Andrew Y, Michael I Jordan, Yair Weiss, et al. (2002). "On spectral clustering: Analysis and an algorithm". In: *Advances in neural information processing systems* 2, pp. 849–856.

Niepert, Mathias, Mohamed Ahmed, and Konstantin Kutzkov (2016). "Learning convolutional neural networks for graphs". In: *International Conference on Machine Learning*, pp. 2014–2023.

Palla, Gergely et al. (2005). "Uncovering the overlapping community structure of complex networks in nature and society". In: *Nature* 435.7043, pp. 814–818.

Palomo, Esteban J and Ezequiel López-Rubio (2016). "The growing hierarchical neural gas self-organizing neural network". In: *IEEE transactions on neural networks and learning systems*.

Papadopoulos, Fragkiskos, Rodrigo Aldecoa, and Dmitri Krioukov (2015). "Network geometry inference using common neighbors". In: *Physical Review E* 92.2, p. 022807.

Papadopoulos, Fragkiskos, Constantinos Psomas, and Dmitri Krioukov (2015). "Network mapping by replaying hyperbolic growth". In: *IEEE/ACM Transactions on Networking (TON)* 23.1, pp. 198–211.

Papadopoulos, Fragkiskos et al. (2010). "Greedy forwarding in dynamic scale-free networks embedded in hyperbolic metric spaces". In: *INFOCOM, 2010 Proceedings IEEE*. IEEE, pp. 1–9.

Papadopoulos, Fragkiskos et al. (2011). "Popularity versus similarity in growing networks". In: *arXiv preprint arXiv:1106.0286*.

Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena (2014). "Deepwalk: Online learning of social representations". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 701–710.

Pons, Pascal and Matthieu Latapy (2011). "Post-processing hierarchical community structures: Quality improvements and multi-scale view". In: *Theoretical Computer Science* 412.8, pp. 892–900.

Psorakis, Ioannis et al. (2011). "Overlapping community detection using bayesian non-negative matrix factorization". In: *Physical Review E* 83.6, p. 066114.

Radicchi, Filippo et al. (2004). "Defining and identifying communities in networks". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.9, pp. 2658–2663.

Rosvall, Martin and Carl T Bergstrom (2007). "An information-theoretic framework for resolving community structure in complex networks". In: *Proceedings of the National Academy of Sciences* 104.18, pp. 7327–7331.

— (2008). "Maps of random walks on complex networks reveal community structure". In: *Proceedings of the National Academy of Sciences* 105.4, pp. 1118–1123.

Shen, Huawei et al. (2009). "Detect overlapping and hierarchical community structure in networks". In: *Physica A: Statistical Mechanics and its Applications* 388.8, pp. 1706–1712.

Shi, Jianbo and Jitendra Malik (2000). "Normalized cuts and image segmentation". In: *IEEE Transactions on pattern analysis and machine intelligence* 22.8, pp. 888–905.

Tang, Jian et al. (2015). "Line: Large-scale information network embedding". In: *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 1067–1077.

Tenenbaum, Joshua B, Vin De Silva, and John C Langford (2000). "A global geometric framework for nonlinear dimensionality reduction". In: *science* 290.5500, pp. 2319–2323.

Thomas, Josephine Maria et al. (2016). "Machine learning meets network science: dimensionality reduction for fast and efficient embedding of networks in the hyperbolic space". In: *arXiv preprint arXiv:1602.06522*.

Van Dongen, Stijn Marinus (2001). "Graph clustering by flow simulation". PhD thesis.

Vandin, Fabio, Eli Upfal, and Benjamin J Raphael (2011). "Algorithms for detecting significantly mutated pathways in cancer". In: *Journal of Computational Biology* 18.3, pp. 507–522.

Vogel, Christine and Edward M Marcotte (2012). "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses". In: *Nature reviews. Genetics* 13.4, p. 227.

Von Luxburg, Ulrike (2007). "A tutorial on spectral clustering". In: *Statistics and computing* 17.4, pp. 395–416.

Watts, Duncan J and Steven H Strogatz (1998). "Collective dynamics of 'small-world'networks". In: *nature* 393.6684, pp. 440–442.

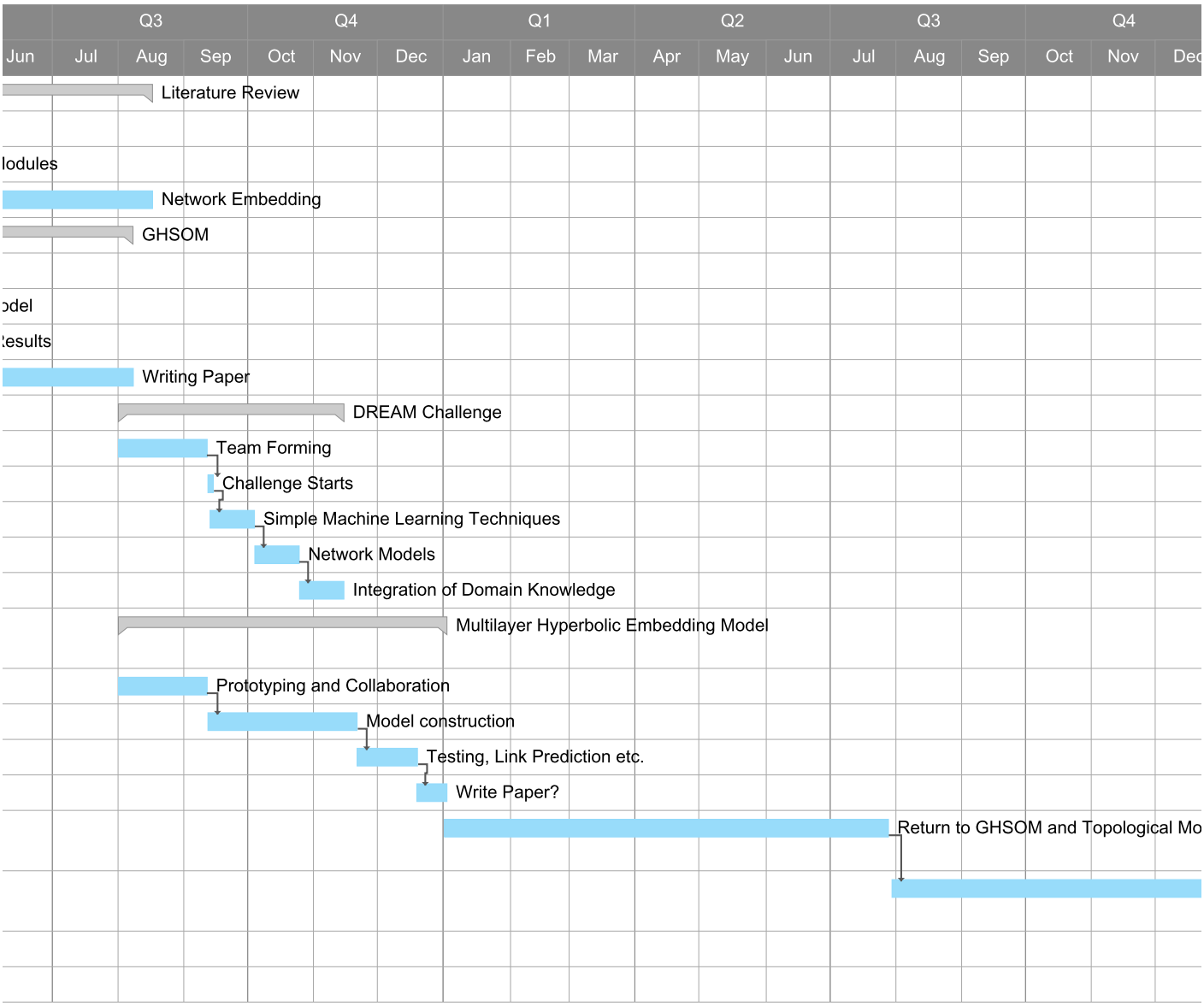
Wei, Y-C and C-K Cheng (1991). "Ratio cut partitioning for hierarchical designs". In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 10.7, pp. 911–921.

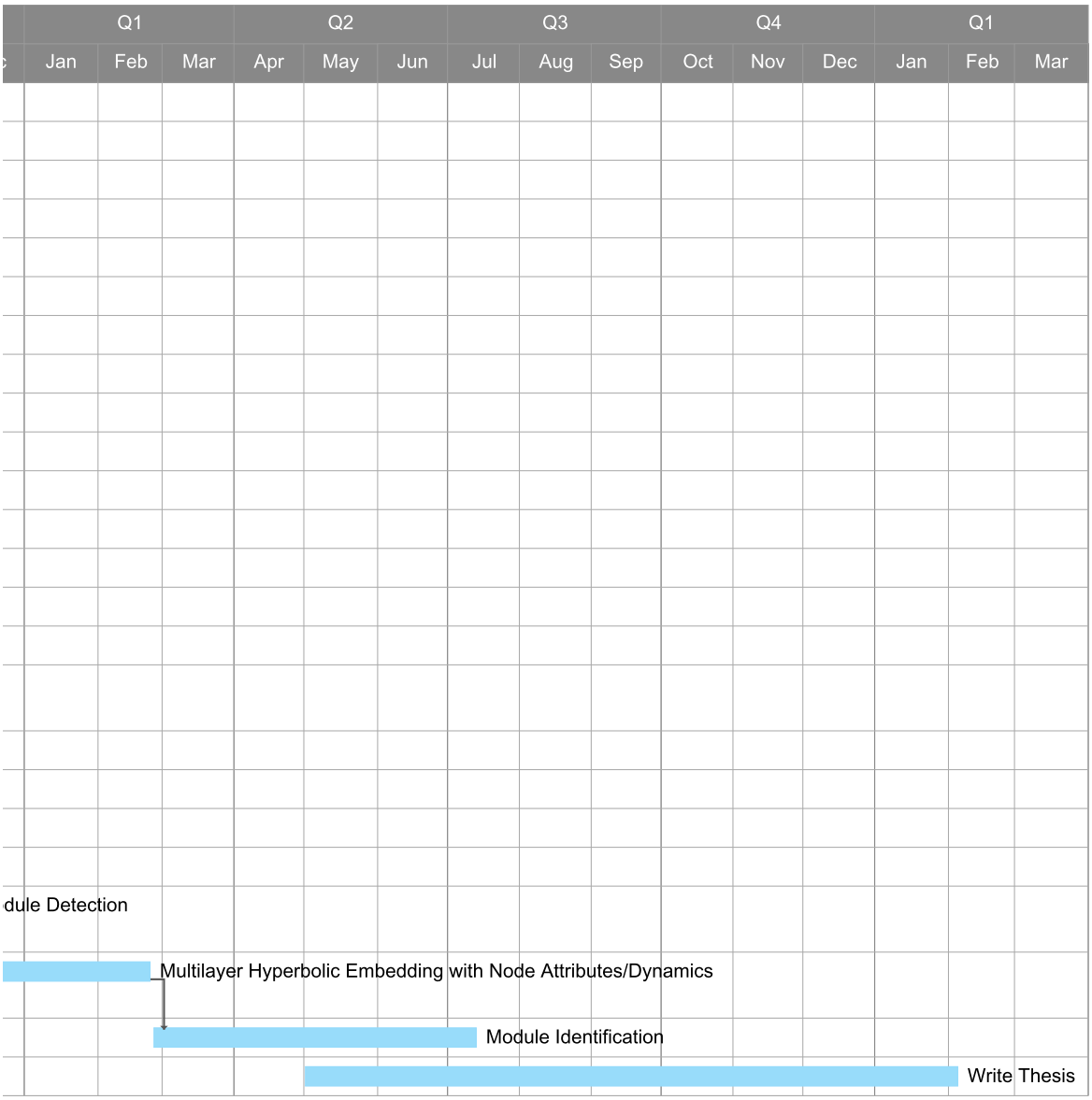
Yamakawa, Takeshi, Keiichi Horio, and Masaharu Hoshino (2006). "Self-organizing map with input data represented as graph". In: *International Conference on Neural Information Processing*. Springer, pp. 907–914.

Zhang, Hui et al. (2016). "Integrated proteogenomic characterization of human high-grade serous ovarian cancer". In: *Cell* 166.3, pp. 755–765.

Basic Project with Gantt Chart

	Task Name	Q3			Q4			Q1			Q2	
		Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
1	Literature Review											
2	Community Detection											
3	Active Modules											
4	Network Embedding											
5	GHSOM											
6	Reading											
7	Construct Model											
8	Results											
9	Writing Paper											
10	DREAM Challenge											
11	Team Forming											
12	Challenge Starts											
13	Simple Machine Learning Techniques											
14	Network Models											
15	Integration of Domain Knowledge											
16	Multilayer Hyperbolic Embedding Model											
17	Prototyping and Collaboration											
18	Model construction											
19	Testing, Link Prediction etc.											
20	Write Paper?											
21	Return to GHSOM and Topological Module Detection											
22	Multilayer Hyperbolic Embedding with Node Attributes/Dynamics											
23	Module Identification											
24	Write Thesis											


Q3
Q4
Q1
Q2
Q3
Q4
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Literature Review
lodules
Network Embedding
GHSOM
odel
esults
Writing Paper
DREAM Challenge
Team Forming
Challenge Starts
Simple Machine Learning Techniques
Network Models
Integration of Domain Knowledge
Multilayer Hyperbolic Embedding Model
Prototyping and Collaboration
Model construction
Testing, Link Prediction etc.
Write Paper?
Return to GHSOM and Topological Mo


Q1
Q2
Q3
Q4
Q1
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
dule Detection
Multilayer Hyperbolic Embedding with Node Attributes/Dynamics
Module Identification
Write Thesis