Extracting Knowledge from Complex Networks: Thesis Proposal

David McDonald
Dr. Shan He
Dr. Peter Tino

Abstract

content...

Contents

Chapter 1

Introduction

sparse E order N small world phenomenon degree distribution follows power law [2, 3] heterogeneity

A complex network is a graph that is comprised of non-trivial or uniform features. These networks often arise when modelling real-word systems. The early belief was that the interactions of such seemingly unrelated things as proteins, social interactions and the internet were random and unconnected. However, it has been shown that most real world systems have the same basic architecture [1]. Real world networks often scale-free in that the distribution of node degree is a power law. Connections are preferentially made between nodes with a probability proportional to their existing degree. Some 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 [2]. The underlying similarity of the interactions between agents in real-world phenomena is surprising and launched the popularity of complex network research. Research that has swiftly captured the imagination of researchers of many fields; fields such as epidemiology, mathematics, computer science, sociology and biology.

Chapter 2

Thesis Statement

2.1 Research questions

1.

2.2 Significance

Chapter 3

Background / Literature Review

Summarise main work in each area

3.1 Community Detection

Most real-world networks contain subsets of nodes that contain a higher degree of inter-connectivity than the rest of the network [3, 4]. These subsets are commonly referred to as communities. A rigorous definition for a 'community' within a network still seems to elude the scientific community [4]. 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 [5] 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 [5]. 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' [6].

3.1.1 Cut-Based and Spectral Approaches

The flagship Kernighan-Lin algorithm [7] 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 [8], normalised cut [9] and min-max cut [10] 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:

$$\phi(S) = \frac{c_s}{\min(Vol(S), Vol(V \setminus S)} \tag{3.1}$$

with

$$c_s = |\{(u, v) : u \in S, v \notin S\}| \tag{3.2}$$

Conductance is still prolific in the literature: it has been used to detect communities in bipartite networks [11], combined with PageRank [12] and used as the basis for a greedy optimisation algorithm [13] capable of finding overlapping communities at different scales.

Spectral clustering dates back to the work of Donath and Hoffman in 1973 [14]. However, it was popularized in the early 2000s [9, 15, 16]. 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 [17], '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.

Further work includes the Markov Clustering algorithm (MCL) that simulates a diffusion process on a graph by repeatedly performing stages of expansion and inflation and only keeping the k largest elements for efficiency [18].

3.1.2 Modularity-Based Approaches

The seminal work of Girvan and Newman [5] 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

$$Q = \sum_{s=1}^{m} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right] \tag{3.3}$$

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 [19], iteratively adding links to a module based on their expected increase in modularity [20], and multi-stage local optimization [21].

3.1.3 Neural Network Approaches

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 [22]. CNNs have also been used for semi-supervised learning on graphs, where the they are capable of learning both graph structure and node features [23].

3.2 Hierarchical Community Detection

The hierarchical nature of modularity-based clustering methods can allow them to detect communities at different scales. [4] 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) [24], and Bayesian non-negative matrix factorisation that performs 'soft-partitioning' and assigns node participation scores to modules [25].

3.2.1 Community Detection in Multilayer and Multislice Networks

3.3 Active Module Identification

Community detection considers only the structure of the network at hand. However, in the age of big data, entities in the 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). [26] 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 [27] in 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:

$$z_i = \Phi^{-1}(1 - p_i) \tag{3.4}$$

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

$$z_A = \frac{1}{\sqrt{k}} \sum_{i \in A} z_i \tag{3.5}$$

where a high z_A represents a biologically active sub-network. Here, Φ^{-1} is the inverse normal CDF.

Computing an exact solution is NP-hard [27], 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) [28], greedy methods [29] and propagation of flow from cancer genes have since been used [30]. More recent work has employed a memetic algorithm to ensure connectedness [31]; a multi-objective optimisation process to control the trade off between biological activity and functional enrichment of the detected modules [32]; and a cooperative co-evolutionary approach [33]. Interestingly, despite the NP-hardness of the problem, [34] 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.

3.4 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 of nodes within the space, such that the probability of reconstructing the observed network is maximised. 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 [35] and Laplacian Eigenmaps [36], 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 [37, 38]. 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 [39, 40]. By introducing additional parameters into the random walk to control a breadth vs. depth first neighbourhood search, [41] were able to identify neighbourhoods of nodes with high *homophily* and high structural similarity.

3.4.1 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 [42, 43]. 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 [44].

The most popular of all these models is the Popularity-Similarity (or PS) model [45]. 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 [46]. This model has a simple interpretation in two dimensional hyperbolic space, where 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[1]:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \text{arccosh}(\cosh(r1)\cosh(r2) - \sinh(r1)\sinh(r2)\cos(\Delta\theta) \tag{3.6}$$

$$\Delta\theta = \pi - |\pi - |\theta_1 - \theta_2|| \tag{3.7}$$

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

Maximum likelihood (ML) was used in [45] 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 in [47, 47]. Due to the computationally demanding task of maximum likelihood estimation, often heuristic methods are used. For example, [46] 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 [48]. Additionally, [49] propose the use of classical manifold learning techniques in the PS model setting with a framework that they call *coalescent embedding*.

3.4.2 Embedding with Node Dynamics

As discussed earlier, it is informative to study not just the topological features of a network, but also the dynamics that take place upon it. NOTHING WITH NODE SCORES

[1] This is the hyperbolic law of cosines.

Chapter 4

Proposed Work / Methods

GAPS:

clustering (attribute based) communities (topology based) (PPI communities rarely align with functional modules) small networks / scalability

(hyperbolic) Embedding with node scores Active Modules in multilayer networks Hyperbolic embedding of multilayer networks Another dimension of attractiveness

4.1 Mathematical Model?

Chapter 5

Preliminary Results

Chapter 6

Work Plan / Timeline

Chapter 7

Implications of Research

Bibliography

[1] Albert-László Barabási. Scale-free networks: a decade and beyond. *science*, 325(5939):412–413, 2009.

[2] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998.

[3] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[4] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

[5] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[6] Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

[7] Brian W Kernighan and Shen Lin. An efficient heuristic procedure for partitioning graphs. *Bell system technical journal*, 49(2):291–307, 1970.

[8] Y-C Wei and C-K Cheng. Ratio cut partitioning for hierarchical designs. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 10(7):911–921, 1991.

[9] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[10] Chris HQ Ding, Xiaofeng He, Hongyuan Zha, Ming Gu, and Horst D Simon. A min-max cut algorithm for graph partitioning and data clustering. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 107–114. IEEE, 2001.

[11] Michael J Barber. Modularity and community detection in bipartite networks. *Physical Review E*, 76(6):066102, 2007.

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

[13] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.

[14] William E Donath and Alan J Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5):420–425, 1973.

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

[16] Chris Ding. A tutorial on spectral clustering. In *Talk presented at ICML.*, 2004.

[17] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[18] Stijn Marinus Van Dongen. *Graph clustering by flow simulation.* PhD thesis, 2001.

[19] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663, 2004.

[20] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[21] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[22] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *arXiv preprint arXiv:1606.09375*, 2016.

[23] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[24] Pascal Pons and Matthieu Latapy. Post-processing hierarchical community structures: Quality improvements and multi-scale view. *Theoretical Computer Science*, 412(8):892–900, 2011.

[25] Ioannis Psorakis, Stephen Roberts, Mark Ebden, and Ben Sheldon. Overlapping community detection using bayesian non-negative matrix factorization. *Physical Review E*, 83(6):066114, 2011.

[26] Koyel Mitra, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. Integrative approaches for finding modular structure in biological networks. *Nature reviews. Genetics*, 14(10):719, 2013.

[27] Trey Ideker, Owen Ozier, Benno Schwikowski, and Andrew F Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(suppl_1):S233–S240, 2002.

[28] Martin Klammer, Klaus Godl, Andreas Tebbe, and Christoph Schaab. Identifying differentially regulated subnetworks from phosphoproteomic data. *BMC Bioinformatics*, 11(1):351, Jun 2010.

[29] Şerban Nacu, Rebecca Critchley-Thorne, Peter Lee, and Susan Holmes. Gene expression network analysis and applications to immunology. *Bioinformatics*, 23(7):850–858, 2007.

[30] Fabio Vandin, Eli Upfal, and Benjamin J Raphael. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, 18(3):507–522, 2011.

[31] Dong Li, Zhisong Pan, Guyu Hu, Zexuan Zhu, and Shan He. Active module identification in intracellular networks using a memetic algorithm with a new binary decoding scheme. *BMC genomics*, 18(2):209, 2017.

[32] Weiqi Chen, Jing Liu, and Shan He. Prior knowledge guided active modules identification: an integrated multi-objective approach. *BMC systems biology*, 11(2):8, 2017.

[33] Shan He, Guanbo Jia, Zexuan Zhu, Daniel A Tennant, Qiang Huang, Ke Tang, Jing Liu, Mirco Musolesi, John K Heath, and Xin Yao. Cooperative co-evolutionary module identification with application to cancer disease module discovery. *IEEE Transactions on Evolutionary Computation*, 20(6):874–891, 2016.

[34] Marcus T Dittrich, Gunnar W Klau, Andreas Rosenwald, Thomas Dandekar, and Tobias Müller. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. *Bioinformatics*, 24(13):i223–i231, 2008.

[35] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[36] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*, pages 585–591, 2002.

[37] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[38] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[39] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077. International World Wide Web Conferences Steering Committee, 2015.

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

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

[42] Dmitri Krioukov, Fragkiskos Papadopoulos, Amin Vahdat, and Marián Boguñá. Curvature and temperature of complex networks. *Physical Review E*, 80(3):035101, 2009.

[43] Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguná. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.

[44] Fragkiskos Papadopoulos, Dmitri Krioukov, Marián Boguñá, and Amin Vahdat. Greedy forwarding in dynamic scale-free networks embedded in hyperbolic metric spaces. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. IEEE, 2010.

[45] Fragkiskos Papadopoulos, Maksim Kitsak, M Serrano, Marián Boguná, and Dmitri Krioukov. Popularity versus similarity in growing networks. *arXiv preprint arXiv:1106.0286*, 2011.

[46] Gregorio Alanis-Lobato, Pablo Mier, and Miguel A Andrade-Navarro. Efficient embedding of complex networks to hyperbolic space via their laplacian. *Scientific Reports*, 6, 2016.

[47] Fragkiskos Papadopoulos, Rodrigo Aldecoa, and Dmitri Krioukov. Network geometry inference using common neighbors. *Physical Review E*, 92(2):022807, 2015.

[48] Gregorio Alanis-Lobato, Pablo Mier, and Miguel A Andrade-Navarro. Manifold learning and maximum likelihood estimation for hyperbolic network embedding. *Applied Network Science*, 1(1):10, 2016.

[49] Josephine Maria Thomas, Alessandro Muscoloni, Sara Ciucci, Ginestra Bianconi, and Carlo Vittorio Cannistraci. Machine learning meets network science: dimensionality reduction for fast and efficient embedding of networks in the hyperbolic space. *arXiv preprint arXiv:1602.06522*, 2016.