Exploring Network Differences between Different Tumour Vessel Networks: Motif Analysis and More

Magdy Saleh
mksaleh@stanford.edu

Surag Nair
surag@stanford.edu

Abhijeet Shenoi
ashenoi@stanford.edu

Abstract

There has been rapid growth in computational modelling within cancer research, particularly investigating the vasculature. New types of datasets have arisen that allow novel analysis of tumour physiology. For this project we propose utilizing a graph based dataset of the tumour blood vessel network of two different tumour types (LS174T and SW1222). The aim is to characterize key differences in the structure and properties of these tumour types, as well as compare their similarities with different networks (random networks as well as other biological and social networks). We plan to explore motif detection and graphlet analysis on the networks, as well as attempt to identify communities and recurring roles. In doing so, we hope to derive insights that can be tied back to the relevant physiology of these tumours.

1 Paper Reviews

1.1 Community detection

Within large networks, communities arise that are characterised by many edges between the nodes of the same community, and few edges between different communities. Community detection methods include different algorithms that allow us to identify a partitioning of nodes that maximize some objective, as described above. Below we review two papers. The first addresses generic methods for community detection based on spectral-modularity based approach and a recursive algorithm. The second paper discusses applying a community detection method on the web graph, where a series of seed nodes are used as starting points and iteratively expanding to capture larger and larger communities.

1.1.1 Modularity and community structure in networks

Summary: Newman [2006] presents a spectral approach to community detection in graphs. Communities are found mainly by optimizing modularity, a quantity defined by

$$Q = \frac{1}{4m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) s_i s_j. \tag{1}$$

Essentially, a community structure is discovered in a graph when you find more than the expected number of edges between a group of nodes.
This function can be rewritten in matrix form as

$$Q = \frac{1}{4m} s^T B s \tag{2}$$

where $s_i = 1$ if node i belongs to group 1, and $s_i = -1$ if node i belongs to group 2. The matrix B is symmetric as is given by $B_{ij} = A_{ij} - \frac{k_i k_j}{1m}$. The Eigen vector with the largest Eigen value gives a partitioning of the nodes which approximately maximizes the modularity. The author then

moves on to outline how this method can be applied to recursively split the graph into multiple communities. Given an initial split, all further subdivision occurs by calculating changes in modularity, only occuring if this change is positive. Further, the paper uses a technique reminiscent of the Kernighan-Lin algorithm to repartition the nodes and further increase the modularity of the resulting partition.

Critique: This work, although seminal, has the following shortcomings:

- Only detects non-overlapping communities
- The quantity being optimized (modularity) only takes into account the edges between nodes in a community
- The author also does not allude to motifs and graphlets or any substructure within the graph.

Brainstorming: One possible improvement to this work improving on the null model, to not place edges at random, but instead to mimic the motif/graphlet distribution within the graph.
Another possible extension is to detect overlapping communities, such as in Xie et al. [2013]. This is particularly relevant to the work we plan on doing, because networks within tumour cells are likely to have multiple overlapping communities. Further, modifying the objective to include the modularity and take into account node characteristics might lead to a better partitioning of nodes, similar to work done in Yang et al. [2013], which takes into account node attributes.

1.1.2 Self-Organization and Identification of Web Communities

Summary: In the paper Flake et al. [2002] the self organization of the web graph is discussed. In particular the work analyzes how communities in the web graph can be directly identified by looking purely at connectivity methods.

The work defines a community as being a collection of nodes in a graph where each node has more edges within the community than outside. In order to recast this NP-hard problem of detecting such communities, the paper assumes one more seed nodes in the graph. Then starting from said seed nodes a Max Flow-Min Cut approach is used in order to identify different communities, based on different thresholds for the fraction of edges required to be within the community for each node. (The standard definition assumes this threshold to be 50%).

Two main algorithms are described in the paper. The first is *Exact-Flow-Community Algorithm.* Given some $G = (V, E)$, and an initial seed set $S \subseteq V$ an artificial source node is added and connected to all $v \in S$, with an infinite edge capacity. Then it makes each existing edge bidirectional and re-scales them and then routes all other nodes to the source node with unit capacity. Then a max flow procedure is called to produce a residual flow graph. The community then becomes the set of all nodes, that are accessible from the source term through positive edges.

The second algorithm specified is the *Approximate-Flow-Community Algorithm* is given some seed set $X \subseteq V$ and crawls to some fixed depth through both inbound and outbound edges to the seed nodes then applies the *Exact-Flow-Community Algorithm* and adds the highest ranked non-seed site to the seed set and then reruns the algorithm for the desired number of loops.

Critique: Since these methods rely purely on node connectivity, as opposed to text processing based methods, means that this can be used on many other generic graph problems. Specifically this can be applied to the tumour networks in order to identify communities of vessels.
The methods outlined in the paper allow for polynomial time computation of communities. The main point of concern is relying on specific previously chosen seed nodes to start the algorithm. These seem to really impact the outcome of the algorithm and play an important role. These make sense if there exits some common theme or relation between nodes that are already known and can be exploited, however it does hinder these methods to be applied without much knowledge of the underlying structure of the graph and so lessens the generality of these algorithms. One problem is that given our cancer dataset, our choice of seed nodes might mean that we already predetermine our results to things already known.

Brainstorming: One way to extend on the work of the paper is to try eliminate the need for heuristically chosen seed nodes. This can be done by iteratively applying this algorithm on randomly chosen seed nodes and check if there are clear communities that arise repeatedly. Secondly, choosing

the seed nodes as peripheral nodes would allow us to identify if there are strong communities at the edges and if so, determine whether the flow (of information, blood, or anything else) really penetrates through the network or only remains concentrated around the peripheries.

1.2 Motif Discovery

Motif discovery in networks captures certain subgraphs of the network that appear multiple times in the graph. Recurring motifs in graphs often highlight certain functional properties that characterise the network. However, discovery of motifs is computationally challenging. We review a paper that proposes an enhanced algorithm for detection of motifs.

1.2.1 MODA: An efficient algorithm for network motif discovery in biological networks

Summary: In Omidi et al. [2009] the authors present an extension of the motif discovery algorithm using subgraph enumeration and symmetry-breaking by Grochow and Kellis [2007]. Specifically, they extend the method to count motifs in induced and non-induced subgraphs (subgraphs with a subset of edges). The method works by building an expansion tree of motifs. Each node of this tree is a motif with k nodes. At each level an additional edge is added to the parent. At level 1, the expansion tree consists of all motifs with k nodes and $k-1$ edges, i.e. all tree with k nodes. The level $\frac{k^2-3k+4}{2}$ has only one node- the complete graph with k nodes. The search for motifs with k nodes in a given network begins by looking for all instances of motifs with k nodes and $k-1$ edges in the network and then storing them. The algorithm then iterates over nodes in the expansion tree. At each node, it retrieves instances of its parent's motifs in the network and tests if an edge can be added to each instance to construct its own motif, in which case it stores the subgraph as an instance of the motif. It also utilizes symmetry-breaking inspired by Grochow and Kellis [2007] to prevent counting the same motif multiple times. They show performance improvements over existing methods for motif counting.

Critique: The paper is an improvement over Grochow and Kellis [2007] and extends the methods to count motif in non-induced subgraphs in a clever fashion. The utility of counting motifs in non-induced subgraphs becomes more important when the motifs have a large number of nodes ($\geq$8). This is especially true in biological networks where there may be spurious or noisy edges between nodes that may contaminate otherwise strong motifs. Counting motifs in non-induced subgraphs will then reveal motifs that may be missed by other methods that count motifs only on induced subgraphs. While the presentation of the algorithm was intuitive and well motivated, the evaluation section was particularly lacking and did not describe the exact setup of the experiments.

Brainstorming: The authors also compare their algorithm to FANMOD (Wernicke and Rasche [2006]) which performs exact subgraph enumeration and performs better in some cases. However, the FANMOD algorithm only counts motifs in induced subgraphs. The authors did not perform extensive experimentation or comment about the performance with respect to FANMOD. We feel that FANMOD can be modified to count motifs in non-induced subgraphs. For each motif found at the end of exact subgraph enumeration, edges can be recursively deleted to count motifs in non-induced subgraphs. A closer analysis is required to validate the difference in performance between the algorithms.

Faster algorithms for motif discovery are a promising avenue for further research. Borrowing from the ideas in Omidi et al. [2009], we think that it should be possible to extend the pattern growth approach to systematically identify larger motifs. Analogous to the expansion tree method that grows the motif one edge at a time, it may be possible to expand the motif one node at a time. Doing so will enable discovery of larger motifs by reusing information from discovery of smaller motifs.

2 Proposal

2.1 Introduction

Small tumours survive by drawing on oxygen and nutrients supplied from the nearest existing vasculature. Once a tumours size exceeds 1-2 mm^3, it can no longer survive based on the diffusion limit of oxygen and nutrients. In order to continue growing, the tumour and starts an angiogenic process to build its own vasculature to sustain its supply of oxygen and nutrients. tumour vasculature

consists of vessels recruited from pre-existing networks as well as angiogenic vessels and differs significantly from healthy vasculature on both a micro (vessel properties) and a macroscale (network geometry).

There has been some work done into analyzing tumour vessels from a network perspective. Skinner et al. [1990] reported higher frequency of branching within tumour vasculature compared to healthy vasculature. These findings were expanded by Less et al. [1991] who directly studies the branching patterns of tumour vasculature and found that two different types of branching patterns. The first was characterized by decreasing vessel diameter and length in successive generations of vessels. The second was characterized by fluctuations in both radius and length across higher degrees of branching. They reported the presence of loops in the vasculature and categorized them into "self-loops", which are loops between two nodes consisting of just two edges and "true loops" which are loops consisting of multiple nodes. Pries et al. [2010] suggested based on the above properties that this can give rise to a "shunt problem" in the tumour vasculature, whereby low resistance, short paths divert blood away from longer paths.

2.2 Data Set

From the Walker-Samuel Lab at UCL, we obtained two different networks for tumours for colorectal adenocarcinomas. The first network is that of an LS174T mouse model of a colorectal carcinoma. This network has $\sim 18k$ nodes and $\sim 26k$ edges. The second network is that of an SW1222 carcinoma, which contains $\sim 72k$ nodes and $\sim 108k$ edges. These two tumour models have markedly different phenotypes as described in D'Esposito et al. [2015]: SW1222 tumours are well differentiated with well-perfused vascular network, whereas LS174T tumours show moderate-to-poor differentiation and are comparatively less-well vascularized and perfused, with larger areas that have died due to lack of oxygen supply. Based on simulations ran by the Walker-Samuel lab, we posses scalar values on each of the edges that describe the simulated pressure, stress and flow.

2.3 Aims:

d'Esposito et al. [2018] , found that based on the two different network structures there are differences in functional connectivity and redundancy in these networks. We plan to expand on this network analysis by considering the following aims:

I. Explore structural difference between SW1222 and LS174T tumours through graphlet and motif analysis

II. Identify the presence, or lack thereof, of recurring roles and different communities / clusters of nodes in the different tumour types with the target of linking that back to the relevant oncological physiology

III. Compare different network structures to different null models to assess the stochastic properties of the models we have

2.4 Methods and Algorithms

Borrowing from the critiqued papers, we plan to analyze the tumours from a functional and structural point of view:

- We will perform a basic analysis of network properties including node degree distribution, clustering coefficient, density and strong and weak connected components for both our networks and highlight the key differences between the two. In addition, we will endeavour to tie the observations back to the biology of the tumours.
- Following Omidi et al. [2009] and Wernicke and Rasche [2006], we will extract motifs in the two different network structures. We are interested in the motifs in both the induced and non-induced subgraphs. We postulate that due to the inherently noisy nature of the biological data, motif discovery on non-induced subgraphs will be more informative than that on induced subgraphs, especially for motifs having many ($\geq$8) nodes.
- We will compare the network significance profiles of the two networks with each other and with other biological networks (e.g. neuronal networks). This will give us a better idea of

the network structure in comparison other biological networks. In addition, we will also compare it with networks such as social networks, language networks and web networks to further pinpoint the exclusive properties of the tumour networks. Once we extract these motifs, we will explore their functional properties in the context of oncological physiology.

- Following Flake et al. [2002] and Guo et al. The goal is to identify communities within these graphs to ascertain clusters of nodes which correspond to specific levels of blood flow. Applying Yang et al. [2013] or a spectral approach such as Guo et al., seems most likely to lead to good results. Also, rewriting the modularity function to take into account node characteristics within communities, might lead to a more semantically meaningful partitioning. Further, trying the algorithm in Blondel et al. [2008], will be used as a computationally cheap way to set up a baseline.
- Additonally, Applying Flake et al. [2002] using peripheral nodes as seed nodes will allow us to determine whether communities are concentrated around the outside of the network and so are routing a lot of the blood flow away from the center of the tumour. This is relevant because it leads to greater oxygen starvation and a higher rate of signalling for more blood vessels to be built, aggravating the growth of the tumour.

2.5 Evaluation and Challenges

Due to the exploratory nature of the project, most of our evaluation criteria will have to ultimately be qualitative. We will have to consider to what extent we have found significant difference or similarities between the different networks and does this fit consistently with existing domain knowledge surrounding those two tumour types. One potential quantitative evaluation metric we can use is to try and predict to what extent the graph structure impacts flow in the network, to which we have simulated values per edges that we will consider ground truth for the sake of this project.

Another challenge we face is that current techniques to evaluate the quality of a partitioning often rely only on the modularity, or require some heuristic for finding partions. This means that we will have to consider the limitations of our chosen heuristics when considering our results and it allows for bias to seep into the analysis. Furthemore, since graph based analysis has not been applied to this dataset or type of dataset before, there does not exist a precedence on how such questions are to be explored in this scenario.

References

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.

Angela D'Esposito, Daniil Nikitichev, Adrien Desjardins, Simon Walker-Samuel, and Mark F Lythgoe. Quantification of light attenuation in optically cleared mouse brains. *Journal of biomedical optics*, 20(8):80503, 8 2015. ISSN 1560-2281. doi: 10.1117/1.JBO.20.8.080503. URL `http://www.ncbi.nlm.nih.gov/pubmed/26277988http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4568291`.

Angela d'Esposito, Paul W. Sweeney, Morium Ali, Magdy Saleh, Rajiv Ramasawmy, Thomas A. Roberts, Giulia Agliardi, Adrien Desjardins, Mark F. Lythgoe, R. Barbara Pedley, Rebecca Shipley, and Simon Walker-Samuel. Computational fluid dynamics with imaging of cleared tissue and of in vivo perfusion predicts drug uptake and treatment responses in tumours. *Nature Biomedical Engineering*, 2(10):773–787, 10 2018. ISSN 2157-846X. doi: 10.1038/s41551-018-0306-y. URL `http://www.nature.com/articles/s41551-018-0306-y`.

Gary William Flake, Steve Lawrence, C Lee Giles, and Frans M Coetzee. Self-organization and identification of web communities. *Computer*, 35(3):66–70, 2002.

Joshua A Grochow and Manolis Kellis. Network motif discovery using subgraph enumeration and symmetry-breaking. In *Annual International Conference on Research in Computational Molecular Biology*, pages 92–106. Springer, 2007.

Yunqi Guo, Xueyin Yu, and Yuanqi Li. Spectral methods for network community detection and graph partitioning.

Joanne R. Less, Thomas C. Skalak, Eva M. Sevick, and Rakesh K. Jain. Microvascular architecture in a mammary carcinoma: Branching patterns and vessel dimensions. *Cancer Res.*, 51(265):265–273, 1991. ISSN 0008-5472.

M Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 2006. ISSN 0027-8424. doi: 10.1073/pnas.0601602103.

Saeed Omidi, Falk Schreiber, and Ali Masoudi-Nejad. Moda: an efficient algorithm for network motif discovery in biological networks. *Genes & genetic systems*, 84(5):385–395, 2009.

A R Pries, Michael Hopfner, Ferdinand le Noble, M. W. Dewhirst, and Timothy W. Secomb. The shunt problem: control of functional shunting in normal and tumour vasculature. *Nat Rev Cancer*, 10(8):587–593, 2010. ISSN 1474-175X. doi: 10.1038/nrc2895. URL `http://www.ncbi.nlm.nih.gov/pubmed/20631803`.

Stewart A Skinner, Peter J M Tutton, and Paul E O'Brien. Microvascular Architecture of Experimental Colon Tumors in the Rat. *Cancer Research*, 50(8):2411 LP – 2417, 4 1990. URL `http://cancerres.aacrjournals.org/content/50/8/2411.abstract`.

Sebastian Wernicke and Florian Rasche. Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, 2006.

Jierui Xie, Stephen Kelley, and Boleslaw K Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, 45(4):43, 2013.

Jaewon Yang, Julian McAuley, and Jure Leskovec. Community detection in networks with node attributes. In *Data Mining (ICDM), 2013 IEEE 13th international conference on*, pages 1151–1156. IEEE, 2013.