Dr. James Girard Summer Undergraduate Research Program  
Faculty Mentor – Project Application

Due Date: *Friday, January 21st at 11:59pm*

Faculty Name: Piotr Szczurek

Department: ECaMS

Research Project Title: Using Link Analysis and Deep Machine Learning for Role Discovery in Networks

Research Project Summary (Please provide an overview of your project -- this will be shared with students as a project description; maximum 500 words):

This project involves analyzing network data such as friendships on social networks, links in transportation networks, or neural connections. This will be done by examining the topological properties in order to discover roles played by particular nodes in the network. For example, in a social network, a particular node may refer to a person and the link between nodes may specify friendship. The node role in such a network could be an influencer or someone how serves as a bridge between communities. There are many reasons for why this would be important. In the social network example, automatic discovery of network roles can serve the purpose of increasing understanding of social structure, providing better marketing information, as well as enabling identification of potentially malicious accounts.

Although the concept of role discovery is not new (sociologists have proposed methods for this), role discovery in Big Data has not been heavily studied. This project is a continuation of previous work that has generated features from network data and used machine learning techniques for classifying network functional roles based on those features. The goal of this work would be to implement most recently proposed approaches for this type of problem, and compare to methods tested in previous studies. Specifically, the current state-of-the-art approach is to use a method called a graph convolutional neural network. This method has been utilized by other researchers to perform role discovery with high accuracy. However, it’s not known whether this method could perform better than using an ensemble model based on link analysis features, which we have tried in previously. We will aim to perform this comparison on a variety of datasets and attempt to further improve upon both with a combination model.

The project would involve programming in Python and performing experiments on different datasets. To get started, the student will have to familiarize themselves with network theory, role discovery algorithms, and Python. They can then proceed to experiment with different methods and attempt to maximize performance on different datasets. Lastly, they would be responsible for documenting the processes and coming up with ideas for trying to improve results.
In the space below, provide a **Project Description** of your research project in 6 pages or less that includes the following:

- **Introduction and Background**
  - Broader impacts on field of study and undergraduate research
  - Specific aims and goals of the project

- **Proposed Research Plan**
  - Preliminary results (if applicable)
  - Experimental design and methodology
  - Instrumentation required and accessibility
  - Expected outcomes and how you will determine success

- **Mentorship Plan**
  - The role of the student in the proposed research plans
  - Plans for engaging with the student throughout the research process
  - How you will hold your student accountable for completing the proposed work
  - Plans for developing specific skills and techniques during mentorship

- **Proposed Timeline that includes Aims and/or Goals**

- **References** (not included in the 6 page limit)

- **Budget up to $500** (not included in the 6 page limit)
  
  *Only include supplies that are absolutely necessary for completing the proposed work (supplies which you do not currently have access to through your department).*
Project Description

I. INTRODUCTION AND BACKGROUND

A network or graph is a set of nodes (or vertexes), connected via links (or edges). Examples of real-life networks include the Internet, social networks, transportation networks, or connectomes (i.e. brain networks). Each node in the network may be associated with certain properties that describe it. For example, in a connectome, some nodes (representing neurons) can have sensory, motor, or interneuron roles. Edges may also have properties. For example, in a transportation network, each edge can be associated with a distance value. Networks may be directed or undirected. In directed networks, each edge connects the nodes in one direction and hence represents an asymmetrical relationship (e.g. one-way streets in a transportation network). In directed networks, all edges represent symmetrical relationships.

The availability of network data has been expanding significantly in recent times. This growth has been spurred by increases in technology which can maintain relationships between entities and record them in real-time. For example, fMRI technology allowed for discovering networks of human brain activities called functional connectomes [1]-[3]. Social networks constantly record all activities by its users and one can analyze the relationships between these activities and the topological properties of those networks [4]. Data gathering by government agencies allowed for creation of crime network datasets, which aids in combating organized crime [5]. All such networks provide significant amount of information and a lot of research has been devoted to knowledge discovery in such datasets. Examples include algorithms for link analysis, community detection, or link prediction [6]-[9].

Very recently, much research has been devoted to the problem of discovering the functional roles of nodes in networks [10]. For example, in a connectome, one may try to discover whether a particular neuron has a sensory, motor, an interneuron role within the brain [13]. In an air-traffic network, role discovery could be used to predict most frequented airports [22]. In a social network, role discovery could be used for finding people with great social influence [24]. It could also be potentially used for finding people with malicious intent. In general, the notion of a functional role of a node is very domain dependent, which makes it difficult to apply a predictive model from one application to another.

Common approaches to role discovery in networks start with so called embeddings. An embedding is a transformation of unstructured data so that the transformed representation makes it easier to solve the given problem. For example, in image analysis, convolutional neural networks are typically employed to generate a feature vector that can succinctly represent various patterns found in the image. This can be viewed as an automatic process of feature generation for a given dataset. These embeddings can be used to make it easier to perform tasks like object detection or image clustering. A network embedding is an embedding based on the set of vertices and edges that comprise the network. Typically, network embedding algorithms will generate a feature vector for each of the nodes in the network, that encode the local and sometimes global topology properties in the vicinity of that node. In the past few years, there have been several algorithms proposed for network embedding. These algorithms include ReFex [26], node2vec [25], and struc2vec [21], many of which can be generalized based on matrix factorization [23]. However, while these network embeddings provide some information about network roles, it’s not clear how they map to roles in a concrete domain. To find these mappings, one can make use of supervised machine learning algorithms, which rely on having a sample of network data with labeled roles for the sample nodes. The algorithms can then generate a model which can predict the roles for nodes outside the sample. The current state-of-the-art approach has been first proposed back in 2017 by Kipf and Welling using Graph Convolutional Networks (GCN) [29]. Since then, it has been shown to achieve a high accuracy on several datasets, although to our knowledge, not on the C.elegans dataset we have used in the past. An alternative to using network embedding algorithms is to use link analysis algorithms such as PageRank [14] or HITS [15] to generate node features, which can then be mapped to specific roles. In our prior work, we have used this successful to achieve a relatively high accuracy on the C.elegans dataset. We have also attempted to use other algorithms for the same purpose, such as node2vec, but discovered that while that has not improved performance by itself, a combined approach can produce much more consistent results. Our goal would be to first compare this approach with that of the GCN method and then attempt to improve the results further by using a combination model approach that we have successfully tried earlier.

A. Specific Aims and Goals of the Project

The basis for this proposed project is the work done in [13] and with subsequent research work done with student projects. In that work, information retrieval algorithms such as PageRank [14] and HITS [15] were used to derive features that, in combination, allow for identifying functional roles of nodes in a connectome. Vectors of such features can be considered as network embeddings that can be used for predictive modeling. Both PageRank and HITS methods were originally used for deriving the importance of web pages on the World Wide Web. In fact, PageRank
was the algorithm that led to the creation of the Google search engine. The *Hyperlink-Induced Topic Search (HITS)* algorithm was designed by Jon Kleinberg as an alternative to PageRank. It provides two measures for each node in a network – a hub score and an authority score. In terms of the Web, a page with a high hub score is one that is linked to by many authoritative pages, while a page with a high authority score is one that is linked to by many significant hub pages. The actual definitions of hub and authority are recursive:

\[
\text{auth}(p) = \sum_{i=1}^{n} \text{hub}(i) \quad \text{hub}(p) = \sum_{i=1}^{n} \text{auth}(i)
\]

, where \( p \) is a web page on the Web, \( \text{auth}(p) \) is its authority score and \( \text{hub}(p) \) is its hub score. Kleinberg showed how these two scores can be derived through an iterative algorithm. In order to guarantee convergence, a normalization invariant that keeps the squared sum of the scores equal to one is maintained in each iteration.

By using the HITS algorithm on connectomes, the results showed we may be able to learn the different roles that neurons can play in brain processing. Prior work showed that the hub and authority scores provided by the HITS algorithm can provide information significantly different from existing methods like PageRank and help to increase the accuracy of role prediction.

Our later work used network embedding algorithms like graphwave [30] and node2vec [25] to attempt to improve the performance of role discovery. The results for node2vec are shown in the table below. The results showed link analysis features provided the highest accuracy, but using a combination model with the node2vec embedding features allowed for much greater consistency of results (lower standard deviation).

<table>
<thead>
<tr>
<th>Fold</th>
<th>Link Analysis</th>
<th>node2vec</th>
<th>Combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>82.76%</td>
<td>65.52%</td>
<td>72.41%</td>
</tr>
<tr>
<td>2</td>
<td>67.86%</td>
<td>60.71%</td>
<td>64.29%</td>
</tr>
<tr>
<td>3</td>
<td>85.71%</td>
<td>67.86%</td>
<td>75.00%</td>
</tr>
<tr>
<td>4</td>
<td>67.86%</td>
<td>64.29%</td>
<td>67.86%</td>
</tr>
<tr>
<td>5</td>
<td>50.00%</td>
<td>35.71%</td>
<td>46.43%</td>
</tr>
<tr>
<td>6</td>
<td>75.00%</td>
<td>67.86%</td>
<td>71.43%</td>
</tr>
<tr>
<td>7</td>
<td>82.14%</td>
<td>53.57%</td>
<td>71.43%</td>
</tr>
<tr>
<td>8</td>
<td>60.71%</td>
<td>57.14%</td>
<td>60.71%</td>
</tr>
<tr>
<td>9</td>
<td>39.29%</td>
<td>50.00%</td>
<td>50.00%</td>
</tr>
<tr>
<td>10</td>
<td>64.29%</td>
<td>85.71%</td>
<td>78.57%</td>
</tr>
<tr>
<td>AVG</td>
<td>67.56%</td>
<td>60.84%</td>
<td>65.81%</td>
</tr>
<tr>
<td>STDEV</td>
<td>14.89%</td>
<td>13.19%</td>
<td><strong>10.60%</strong></td>
</tr>
</tbody>
</table>

The results were similar when using the graphwave algorithm, as shown in the figure, below.

![Figure 1. Comparison of Link Analysis to graphwave Performance on C.elegans Dataset. The results showed a combination or fused model produced the best results.](image)

The goal of this proposed project is to build upon work done in [13] and in prior research done with students to determine whether the results can be improved upon with the use of Graph Convolutional Networks. We would also
perform a comparison of the method of the network embeddings based on PageRank and HITS to the use of node2vec or graphwave methods. The comparison would be done in terms of both accuracy and the run-time efficiency. Lastly, we would examine how the proposed methods perform on network data from different domains.

B. Broader Impacts on Field of Study and Undergraduate Research

Achieving the outcomes of this research has the potential to enhance the ability of data scientists to analyze network data in various domains. If our aim is achieved, we could produce more accurate models of role discovery that can be critically important for a large range of applications. They could help us understand role of individual neurons and their collections in the brain and potentially, generate features for use in diagnosis of diseases [28]. They could also help in identifying malicious nodes in communication or social networks, or in analysis of transportation networks that can lead to improvement of transportation efficiency. For the student, this project would provide an introduction to the fairly recent subfield of network data science and how it can be applied to various domains. It would also provide an opportunity to gain experience in programming with network data, specifically, using Python and the network package. Given the broad applicability of this research and the skillset it entails, I feel it allows for students from various disciplines to learn skills and knowledge they can use for many years in their careers.

II. Proposed Research Plan

A. Preliminary Results

The initial work on using feature-based role discovery in the connectome of the C.elegans worm has been done in [13]. There, we first compared network embedding features based on two major information retrieval algorithms: PageRank and HITS. Our analysis showed that in the network data for the C.elegans connectome, the hubness value produced by the HITS algorithm produced significantly different type of node importance ordering than that of PageRank, while the authority value (also produced by HITS) was somewhat correlated with PageRank output. We have also done this analysis for Erdos-Renyi random graphs for different parameters of the randomness probability. When examining graphs of similar size to that of the C.elegans connectome network, the PageRank values were completely uncorrelated to hubness, while correlation to authority varied, with a single minimum occurring at \( p = 2^{-7} \). This verified that use of both hubness and authority is appropriate for networks with connectivity that is away from the extrema points. An observation during the analysis, that many nodes have both high hubness and authority or low hubness and authority, led to proposing a new single measure of node importance based on the ratio of hubness to authority (H/A ratio). Experiments showed this new feature allows for better discrimination between the three different types of neuron roles. Using supervised machine learning techniques, we have built models using different combinations of the features and showed that using the H/A ratio improved classification performance from 59% to 65%.

In past SURE research programs, further work was done to try to improve the prediction results by examining new network embedding features and experimenting with various kinds of supervised machine learning techniques. We have found that a linear combination of hubness and authority can be used to produce almost 60% accuracy using a single feature. We have also tried adding several network centrality measures and measures derived from PageRank and HITS to maximize the accuracy and finally achieved approximately 81% accuracy with a 10-fold cross-validation. However, since we have been able to show that a single metric based on the sum of hubness and authority scores provides significant information about the functional role of neurons in the connectome, we decided to further pursue this measure to examine what information it provides. In particular, the results showed that for low values, the functional role tended to be sensory; for high values, it tended to be a neuro-muscular junction, and values in between, tended to be interneurons (see figure, below).

The results we obtained indicated it might be possible to find neurons with similar roles simply by clustering the values of the HITS-derived feature. We therefore used unsupervised machine learning techniques to perform clustering and find the optimal number of clusters that occur in the data. The results were four clusters, each of which having the majority of nodes corresponding to a single role. There was one outlier node, which corresponded to the neuromuscular junction and it was in a cluster by itself (this was expected). This led to the conclusion that it might be
possible to determine, with significant accuracy, the role of network nodes by using a single simple measure. Although the specific range of values for this measure might different for networks from different domains, clustering would reveal those with the same role. Thus, this one metric could potentially be used for a variety of applications, such as medical diagnosis, cybersecurity intrusion detection, or finding social media potential influencers. Yet, it was uncertain whether this single metric could also be used in other types of networks. In a prior work, we have further analyzed the properties of the various network embedding features and found that certain features are well correlated to particular function roles in the network. This is illustrated in the figure 3, below, which shows that nodes with large betweenness centrality (between) are typically interneurons, those with large H/A ratio (haratio) are typically sensory neurons, and those with high harmonic mean of hubness and authority are most likely motor neurons.

Figure 3. Values of network embeddings for C.elegans network nodes using three different features.

A similar analysis was then repeated for network data for other domains, including data based on emails sent and industry partnerships. In these cases, the results were not as clear as for the C.elegans data, but it was still visible that certain feature values corresponded to particular roles. This is exemplified in figure 4 below, where high values of the harmonic mean of hubness and authority tended to coincide with the nodes labeled as “content”.

Figure 4. Probability estimates of role labels from the industry-partnership dataset for different values of the harmonic mean of hubness and authority.

However, it may be also the case that the labeled roles are not determined based on network topology and hence no feature set based on network embeddings could be used to correctly predict such labels. While being able to determine conditions under which network data can be used for role discovery is important, finding theoretical results for this research question would be difficult. Instead, the focus of this proposal is on trying to maximize the performance of predictions on the previously used features and compare them with other approaches of network
embedding. We will specifically apply the GCN method to see if this produces significantly better results than in prior approaches and if it can be combined, as we have done with other approaches, to generate even higher accuracy.

B. Experimental Design and Methodology

The project will be divided into two parts. For the first part of the project, we will follow the work in [13] and work done as part of a previous SURE program, which used data provided by [16] to analyze the C. elegans worm connectome in terms of its hub and authority values and generated a predictive model of neuron role. We will use Python along with the networkx package [17] to load the data, model the network structure, and execute algorithms necessary to derive the proposed network embedding features for the C.elegans data and other publicly available datasets, such as those available in [19]. We will then generate network embeddings based on other approaches used in literature, such as node2vec, graphwave, and GCN. Many of these approaches have Python standard implementations already available through the stellargraph package in Python [31]. We do know, by the Ugly Duckling Theorem, that there cannot be a single best set of features for all domains. Therefore, we would do a comparison of the predictive ability of the various features using both supervised and unsupervised machine learning methods on data from various domains and determine which features are more useful for particular domains. In the second part of the project, we would attempt to generate a fused model, using the different embedding approaches and link-analysis features to determine the limit of machine learning on particular datasets, specifically on C.elegans.

C. Expected Outcomes and Success Will be Determined

It is expected that the use of additional features from GCN and other network embedding techniques will allow for an increased accuracy in particular domains. The results will be measured in terms of accuracy, mean (unweighted) recall, and mean (weighted) precision across all role labels. Although accuracy is generally the most important metric, computing recall and precision across the different roles will verify whether a predictive model captures any information about those roles. This cannot be achieved just by accuracy, since guessing the most common role may achieve a high accuracy, especially in highly unbalanced data. The other goal of the project is to determine a simple set of features or even a single feature that has a high predictability in many domains. We will therefore weigh the predictive models against their complexity and try to find an optimal tradeoff between these two criteria.

III. MENTORSHIP PLAN

The student will be responsible for the following:

- Researching the relevant literature and learning about network theory, algorithms for role discovery, and Python coding
- Using network and stellargraph Python packages to run the HITS, PageRank, GCN, and other algorithms
- Collecting network data from various sources
- Loading the data and modeling it as a network
- Performing statistical analysis of the outputs
- Experimenting with various network embedding and link analysis algorithms from existing literature and comparing results.
- Writing up and presenting the results

Students will have to document their work in a written report and meet with me on a regular basis (usually twice a week) to discuss results and plan future work. They will be required to produce specific deliverables for each portion of the project. First, the student will perform a literature survey and summarize the results to show understanding of the existing methods. Second, they will attempt to replicate the results of previous work and present this during a meeting. This should occur within the first two weeks of the program. During this time, they will also be mentored on proper research documentation, performing a literature survey, and given time and advice on how to learn the required skills for further work. This includes coding in Python, the networkx and stellargraph packages, and machine learning methods. Third, they will research and gather datasets that can be used for analysis and comparisons. They will need to describe each dataset and produce network embeddings based on various methods. These results will be documented in a report and presented during a meeting. Furthermore, they will perform experimentations and cross-validation with various machine learning methods and check whether this can improve the performance of predictive models. Lastly, they will attempt to experiment with fusing the methods to try to further improve the results.
IV. TIMELINE AND GOALS


Weeks 3-4: Gather datasets and write code to generate network embeddings.

Weeks 5-6: Experiment with machine learning methods using different network embedding features and compare results.

Weeks 7-8: Apply ensemble and fusion methods to attempt to increase results. Also, experiment with different hyperparameters and methods to further increase performance.

Weeks 9-10: Continued evaluation of results and experimentation. Write final report. Work on presentation of results.

REFERENCES


[17] https://networkx.github.io/


Description of any additional funding you will be using for your proposed research (Doherty Grant, Lasallian Research Grant, External Research Grant, etc.) and how it will be used in this project.

Criteria for student applicants (Please report minimum criteria you will expect from student applicants, such as coursework that must be completed prior to starting work on this project):

Programming experience (esp. Python). Coursework in discrete mathematics, data science, artificial intelligence or machine learning is helpful, but not required.
As a faculty mentor, you will be required to participate as a leader for one of the weekly student seminars. This will be a 60-minute presentation at 9:00 am. Please indicate topics of interest from the themes listed below, or suggest an additional topic, that you might enjoy presenting.

_____ Ethics in Research

_____ Literature Search and Library Resources

_____ Scientific Method and Problem-Solving Skills

_____ Presentation Skills

___X__ Data Analysis and Data Management

_____ Technical Writing

_____ Resume Writing and Marketing YOU

_____ Preparing for Graduate School

_____ Interview Skills

_____ Mock Presentation Supervisor (Practice for Symposium)

_____ Other (Please Describe)

_____________________________________________
The James Girard Summer Undergraduate Research Program (SURE) is designed to support the execution of this proposed project by the faculty mentor and a single undergraduate student. After review of faculty proposals, selected projects will be advertised to Lewis University students, and all interested undergraduates will then be required to apply into the program, denoting the project for which they would like to be considered. Student applications will be reviewed for completeness by the program director and then forwarded to the appropriate faculty mentor for final selection of a candidate. Faculty may submit up to 2 projects for funding through the program. Although faculty mentors may also mentor additional students in the summer not funded through the program, the weekly program events and presentations will be exclusive for students in the program.

By submitting this application, you are agreeing to the following responsibilities of a SURE Faculty Mentor:

- Working closely with your student to ensure a worthwhile educational experience. Regular interactions with your student are an expectation (a minimum of once a week, but more frequently is encouraged). Interaction with other mentors and students is strongly encouraged
- Participating in the welcome and orientation day
- Leading at least one of the weekly workshops for the entire group of participants
- Writing at least one blog related to your area of expertise for the program website
- Participating in the Summer Research Symposium

This application will be reviewed by a faculty panel for acceptance into the program; determination of selected projects will be communicated after review. Project descriptions will then be made available to Lewis University undergraduate students, who can apply to the program and specific projects online via our website. Student applicants will be matched with mentors using a selection process where mentors rank interested students based on their applications and students rank projects based on their interests.

Any questions and all completed applications should be sent to Brittany Stephenson (SURE Director) at bstephenson@lewisu.edu.