2023-04-07

Network Representation Theory in Materials Science and Global Value Chain Analysis

Mats C. Haneberg
Brigham Young University

Network Representation Theory in Materials Science and Global Value Chain Analysis

Mats C. Haneberg

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Zachary Boyd, Chair
Benjamin Webb
Mark Kempton

Department of Mathematics

Brigham Young University

ABSTRACT

Network Representation Theory in Materials Science and Global Value Chain Analysis

Mats C. Haneberg
Department of Mathematics, BYU
Master of Science

This thesis is divided into two distinct chapters. In the first chapter, we apply network representation learning to the field of materials science in order to predict aluminum grain boundaries' properties and locate the most influential atoms and subgraphs within each grain boundary. We create fixed-length representations of the aluminum grain boundaries that successfully capture grain boundary structure and allow us to accurately predict grain boundary energy. We do this through two distinct methods. The first method we use is a graph convolutional neural network, a semi-supervised deep learning algorithm, and the second method is graph2vec, an unsupervised representation learning algorithm.

The second chapter presents our dynamic global value chain network, the combination of the dynamic global supply chain network and the dynamic global strategic alliance network. Our global value chain network provides a level of scope and accessibility not found in any other global value chain network, commercial or academic. Through applications of network theory, we discover business applications that would increase the robustness and resilience of the global value chain. We accomplish this through an analysis of the static, dynamic, and community structure of our global value chain network.

Keywords: network representation learning, network theory, graph convolutional neural network, network community structure, supply chain, strategic alliance

Acknowledgements

I would like to thank my wife, Ashley, for her constant support and cheerful attitude throughout my whole education. I also want to thank her for motivating me to get my master's degree and fulfill the thesis requirement and for believing in me that I could do it. I could not have done this without her.

I want to thank my advisor Dr. Boyd, for teaching me the most influential part of my education at Brigham Young University, how to research and how to run projects. I would like to thank him for his support and advice that I pursue the thesis requirement for the master's degree. Most of all I want to thank him for his time and help throughout this process.

I also want to thank Dr. Webb and Dr. Kempton for their support and guidance as I worked towards my thesis, their help was greatly appreciated.

Last, I want to thank my sister, Marren, for taking the time to proofread my thesis. Her help greatly improved my writing.

Contents

List of Tables

List of Figures

Chapter 1. Aluminum Grain Boundary Energy Prediction and Atom Level Structure Insights Through Network Representation Learning

1.1 Introduction

Grain boundaries (GBs) between metallic crystallites have a large impact on the properties of the metal they create. These include strength, corrosion resistance, and cracking. Being able to control these properties is essential for using metal for many applications. For example, bridges, planes, buildings, or anything that affects our safety cannot be built with metal that does not meet a series of strength and flexibility requirements. If we can understand how atom placement in GBs impacts each property, we can better design and create these metals with new uses and improved safety. This desire has created the field of GB engineering, the study of how to use thermomechanical processing to enhance desired metal properties such as strength [35]. See Section 1.2.2 for an in depth explanation of GBs.

The goals of this chapter are to apply network theory and machine learning to demonstrate how these fields can help us better understand GB properties from the atoms themselves as well as help predict GB properties from a fixed-length encoding.

In order to do this, we have created graphs from 7304 aluminum grain boundaries, discussed further in Section 1.3. We then applied two machine learning methods, a graph convolutional neural network (GCN), and graph2vec to embed each GB in finite dimensions and measure which atoms inside the crystallites are most influential at predicting GB energy, a key GB feature. Graph2vec is an unsupervised representation learning algorithm, while the GCN is a semi-supervised deep learning algorithm. Each of these methods have different strengths and weaknesses, and allow for atom-level insights of GBs. This knowledge provides

explainability to our methods, a key result of this chapter.

Our contributions are the following:

- Create fixed-length representations capturing GB structure through graph2vec.
- Apply a GCN and graph2vec to effectively predict GB energy values while maintaining necessary GB properties such as invariance to permutation, rotation, and perturbation.
- Draw insights into which atoms in each GB influence GB energy the most through our GCN model.
- Apply a graph2vec model to find potentially impactful substructures inside the GB.

1.2 Background

In order to accurately capture GB properties, embedding techniques must be rotation and permutation invariant. Several GB embedding methods that meet these requirements have been proposed in recent years. Here we discuss some of the more influential methods.

1.2.1 Current Industry Methods for Fixed-Length Grain Boundary Representation. One method for creating a fixed-length representation of a GB is Smooth Overlap of Atomic Positions (SOAP). SOAP is a local atomic descriptor, which means it creates a separate embedding for each atom in the GB embedding. It does this through "a local expansion of gaussian smeared atomic density with orthonormal functions based on spherical harmonics and radial basis functions" [11]. In the paper introducing the dataset we employ our methods on, the authors used SOAP to created fixed-length representations of every atom contained inside a GB and then averaged these representations to create an overall GB fixed-length representation. This method performed well in their tests on predicting GB energy values, returning a R^2 score of roughly .95 and an RMSE close to 13. The graph2vec and GCN methods we applied to the data did not perform as well, but the SOAP methods

suffers from a lack of explainability. One of our main goals was to provide a measure of explainability, and that is an area where we outperformed SOAP.

Another method for atomic embeddings is ALIGNN-d, standing for Atomistic Line Graph Neural Network, where the d signifies we include dihedral angles on top of the bond angles included in standard ALIGNN. ALIGNN-d creates a line graph to represent the bonds between atoms and to encode the dihedral angles between them. In a line graph, the nodes represent the edges of the original graph and edges between nodes represent two edges having shared a node in the original graph. Hence this method allows us to embed edge data using node features in the line graph. The method works by using edge-gated graph convolution on the line graph to update features which are then passed to the edge-gated graph convolution of the original graph. In their paper introducing the ALIGNN-d method, the authors note that while the original ALIGNN method is presented as being effective for both periodic and non-periodic atomic structures, they believe this method might not encode periodic structures as well as it does non-periodic structures. Thus in their testing of ALIGNN-d they only applied their method to non-periodic atomic structures [9, 20]. GBs are periodic structures, hence our application of ALIGNN-d was a new experiment for this method and in a comparison of our results of our application of ALIGNN-d to the above described aluminum GB dataset the method did not perform as well as a 2-layer GCN. Due to this method being significantly more complex than our 2-layer GCN and underperforming the GCN, we decided to not dive into the results of ALIGNN-d; however, it does add strength to our results that we were able to outperform a state-of-the art method in the context of aluminum grain boundaries. Another method to be aware of in this field is Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials (NequIP). This method learns interatomic potentials from ab-initio calculations through E(3)-equivariant convolutions, hence is a great method for GB embeddings [4].

1.2.2 Grain Boundary Basics. Atoms, when permitted, align themselves in well-ordered repeating crystalline structures. The repeated crystalline structure depends on the

atom type. For aluminum, this repeated crystalline structure is called face-centered cubic (FCC) structure. FCC structure involves a cube, created by one atom in each corner, and one atom on every face of the cube [32]. Together, these repeating FCC structures form a grain. When two grains collide, they are not able to reorient themselves in order to create one grain. Thus they create a different structure between themselves that does not follow the typical crystalline structure. This different structure is a GB, the structure of which varies widely and depends largely on the angles at which the two crystalline structures collided. The structure of the GB affects the created metal's strength, corrosion resistance, ductility, and many other features [19]. An example GB from the dataset we used in this chapter is seen in Figure 1.1.

Since GBs are two crystalline structures that are not aligned as one grain, the GB itself is less dense than the FCC crystalline structures. This excess space between the crystalline structures creates free energy per unit area, known as the GB energy value. This GB energy value is the GB property we strive to predict in this chapter. Generally, the more misaligned two crystalline structures are, or the more space that is left between them in the GB, the higher the GB energy value will be [19, 37].

GBs consist of what are called coincidence site lattices (CSLs). CSLs have what is called a $\sum$-value. The $\sum$-value is calculated by viewing a GB in two dimensions, removing the dimension that is parallel to the GB. Viewed in this dimension, the two crystalline structures in the GB have some atoms that align, while others do not. We use a group of four neighboring aligning atoms to create a box, and then how many atoms that are within the box, all of which do not align, gives the $\sum$-value. A perfect FCC structure has a $\sum$-value of 1 as when a FCC structure is viewed in two dimensions, all atoms align. Generally, higher $\sum$-values are associated with higher GB energy values [19].

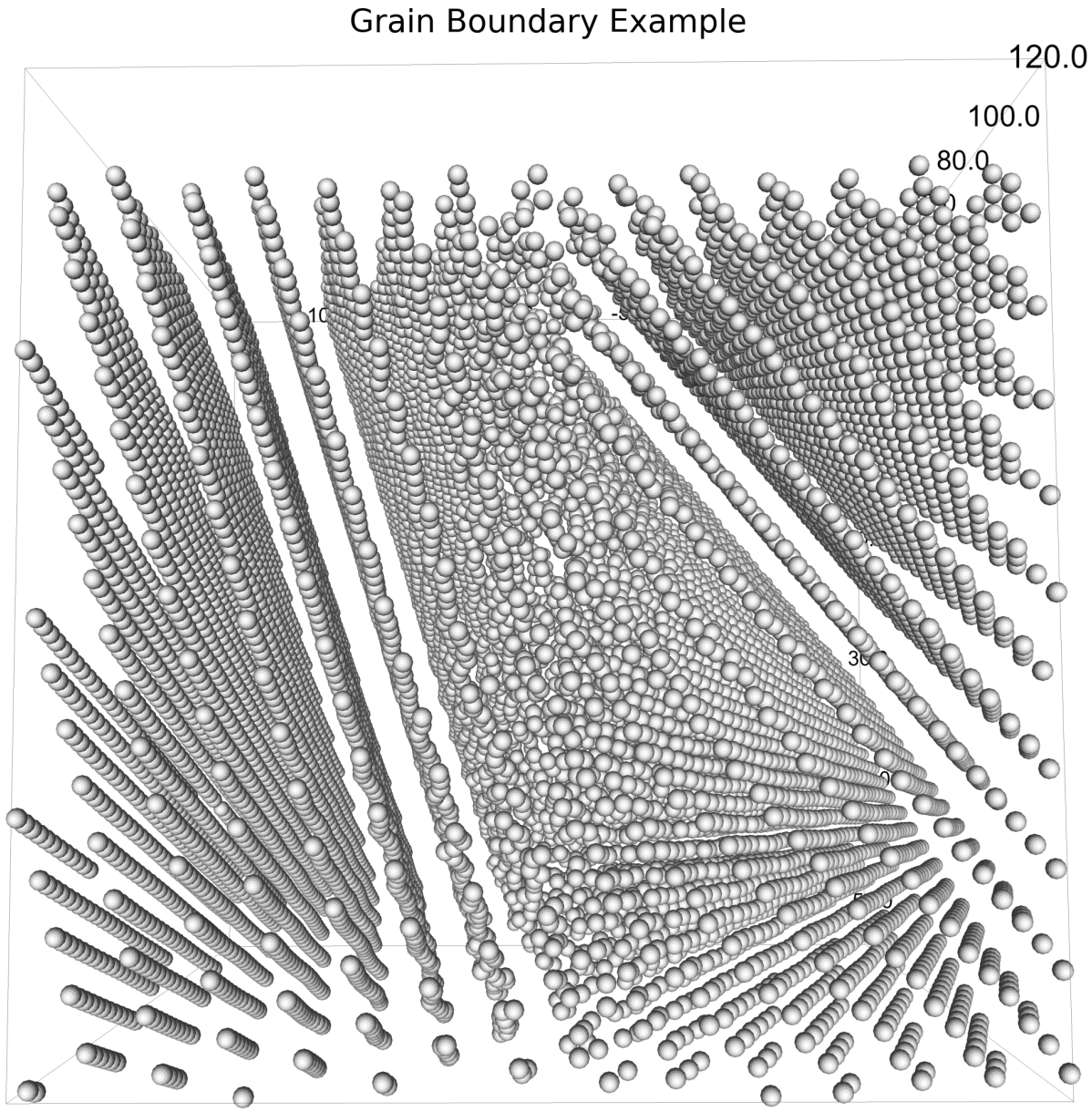

Figure 1.1: GB example from the dataset. The GB is located near the center of the image and travels vertically. Their collision at an angle is visible in the two crystalline structures by viewing the angles of the lattices of atoms on each side, right and left. At the GB's location in the center of the image, we can view how the order of each crystalline structure on either side of the GB breaks down to form the periodic structure of the GB.

1.3 Dataset

We use $7,304$ aluminum grain boundaries in the 5D crystallographic space introduced by Homer et al. in [19]. Since the dataset contained only aluminum atoms, in our graph creation every node in the graph had the same atomic features, rendering these features ineffective for machine learning algorithms, and unimpactful in our research. All grain boundaries in this dataset were drawn from 150 CSLs with $\sum$-values below $1,000$. The authors performed a thorough comparison to ensure each CSL was sufficiently different from the others contained in the dataset to ensure a comprehensive draw of CSLs. They then retrieved minimum-energy grain boundaries from each CSL that helped describe several fundamental zones of the CSL. This process created diverse and descriptive grain boundaries of this space of CSLs with $\sum$-values below $1,000$, making the dataset useful for testing machine learning and embedding methods.

The dataset is stored in Python ASE files, one GB per file [27]. Each file contains the coordinates of the aluminum atoms of the GB as well as a centro-symmetry parameter, indicating how symmetric the positions of neighboring atoms to the central atom are, and a common neighborhood analysis categorical variable, which classifies the type of neighborhood the atom belongs to. These last two variables, the centro-symmetry and common neighborhood analysis variables, are useful in determining whether or not an atom is part of a FCC formation, a standard format for atoms not along or near the intersection of the two crystallites of the GB, which we will refer to as the center of the GB throughout this chapter. Graph2vec only used the positional data to create an embedding, whereas the GCN used all five features to predict GB energy. To create our networks from this dataset, we used the atoms as nodes and then applied epsilon nearest neighbors to connect the nodes, similar to how the authors of ALIGNN-d and NequIP created atomic networks within their code repositories [20, 4]. Grain boundaries are periodic, so we ensured our epsilon nearest neighbors technique took this into account, wrapping the dataset around in the proper dimensions. We tested several different values of epsilon for the epsilon nearest neighbors, and

settled on using 3.5 Ångstrom for graph stability and performance in comparing the GCN and graph2vec algorithms' performance. The process of choosing epsilon is discussed further in Section 1.4.1.2 and Section 1.4.2.2.

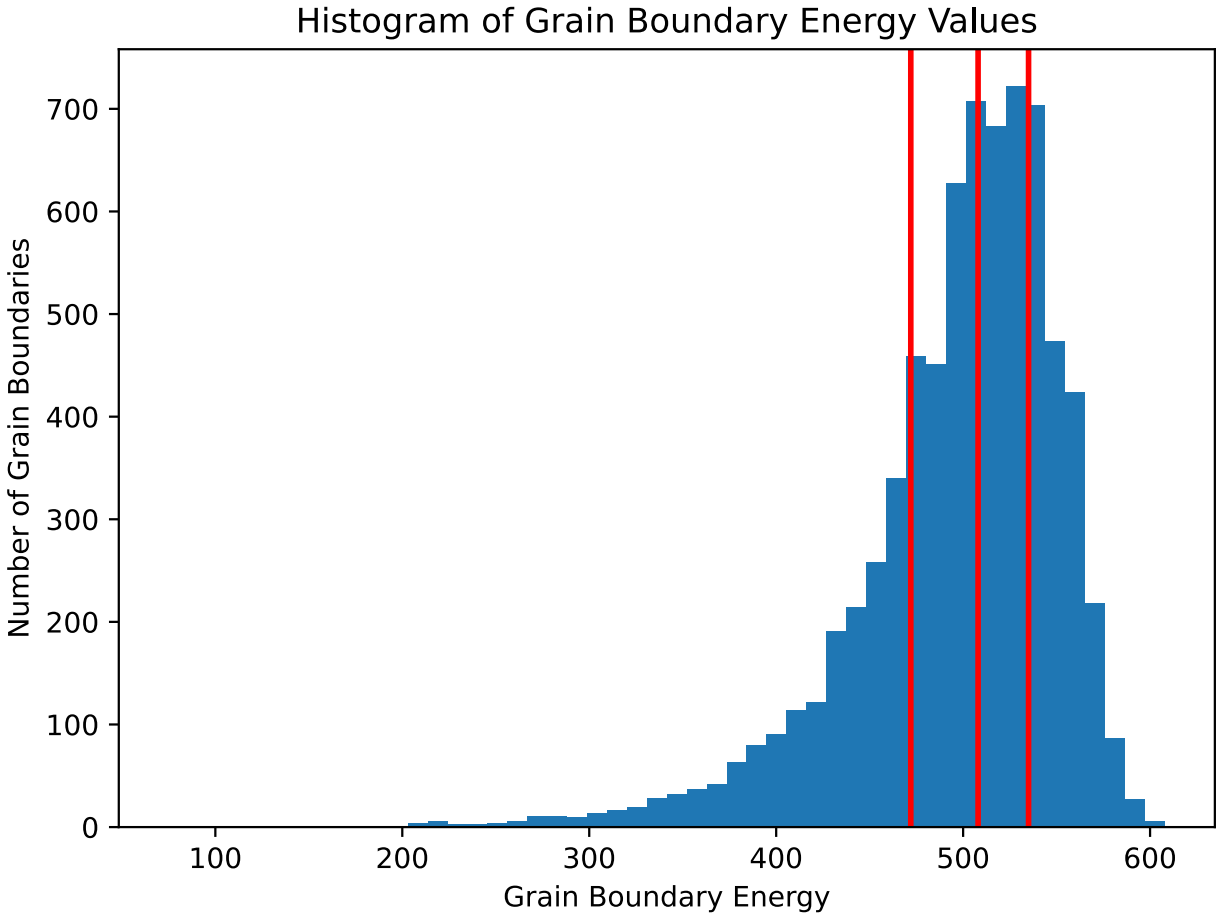

Figure 1.2: Histogram of GB energy values. The vertical lines depict the first, second, and third quartiles of the energy values, at 472, 508, and 535 respectively. The distribution is left-skewed, the lowest 25% of energy values generally fall in a range of 200 to 472, a range much larger than the 473 to 608 range containing the upper 75% of data. It is important to verify our models perform well at predicting GB energy values that fall in the lowest 25% to ensure our models are capable of learning the whole spectrum of GB energy values.

The GB energy values we used for prediction had a left-skewed distribution. The energy values had the following ranges for each quartile, starting at the lowest: 75 to 472, 472 to 508, 508 to 535, and 535 to 608. This distribution is seen in Figure 1.2. The bulk of the bottom distribution was contained between energy values 200 and 472, a range over twice as long as the other three quartiles combined. This discrepancy in the quartile ranges presented

a challenge for our models, as the data was sparse in the bottom quartile's range for learning purposes.

While graph2vec only takes a complete network as input and only trains on one node feature, which we explain in Section 1.4.2, the GCN used all five node features. We tested the GCN in the Cartesian plane, however, this method did not provide the rotation and permutation invariance required. To overcome this we converted all atomic coordinate features to spherical coordinates before training. This produced immediate improvements in performance and was used for all statistics we computed for the GCN. We used a constant random seed to create all train, test, and validation sets, thus results comparing different epsilon values in Ångstrom are being applied to the same data. We used $2,000$ training and validation samples and $1,000$ test samples. Since graph2vec is unsupervised, we did not use the validation data in the model training process. All grain boundaries graphs were created in Python through NetworkX [18].

1.4 Results

We now discuss the findings and perform comparable tests for our GCN and Graph2vec models. Our tests focus on predictive performance and on drawing insights from the atoms in the GB.

1.4.1 Graph Convolutional Neural Network. We next created and applied a GCN to the dataset in order to predict GB energy levels and interpret which atoms were most influential in the energy prediction. We created our GCN through the PyTorch Geometric library in Python [26, 34, 13].

1.4.1.1 How Our GCN Works. We created a 2-layer GCN following the general layer format explained by Kipf and Welling in [26]. GCNs function by reducing the dimension of the input data and by applying features of the neighboring nodes to the base node during convolution. This is similar to how convolutional neural networks (CNNs) when applied to

an image reduce dimension and apply neighboring pixel attributes to a base pixel. Thus by using two convolutional layers, each node's output was affected by its neighbors within a path length of two.

Our first hidden layer had the following format:

$$H^{(1)} = \text{ReLu}(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(0)}W^{(0)}). \tag{1.1}$$

Where $\hat{A} = A + I$ with A the GB adjacency matrix or edge weight matrix if edge weights were used, and $\hat{D}_{ii} = \sum_j \hat{A_{ij}}$ is the diagonal degree matrix of $\hat{A}$. In our testing of various edge weights we found no significant improvement in model performance, and thus left them out of the model for model simplicity. Some weights we tried were Euclidean distance and $e^{-\text{Euclidean distance}^2/\text{radius}}$, which weighted edges closer to the radius near 0. Thus in the case edge weights were not used, $\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}$ is the normalized adjacency matrix of the GB. In our case, $H^{(0)} \in \mathbb{R}^{N\times 5}$ is our input feature matrix where N is the number of atoms in the GB and 5 represents the number of atom features we have; three describing the spherical coordinates of the node, one describing the symmetry of surrounding nodes, and another describing the structure of the neighborhood of the atom. $W^{(0)} \in \mathbb{R}^{5\times 45}$ is our weight matrix and from our testing we found 45 hidden features that produced the best results. Finally, we used a ReLu activation function for the layer.

The second layer of our GCN was formatted the following way:

$$H^{(2)} = \hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(1)}W^{(1)}. \tag{1.2}$$

Where $H^{(1)} \in \mathbb{R}^{N\times 45}$ is as described in 1.1 and $W^{(1)} \in \mathbb{R}^{45\times 1}$ is our weight matrix, for every atom we output one number to represent our prediction for the GB energy.

The last step we applied in our GCN was a global mean pooling layer (GMP) to the outputs of all atoms inside the GB and used this as our GB energy prediction. We trained our model on optimizing the mean squared error and used backpropagation to update the weight matrices $W^{(0)}$ and $W^{(1)}$.

Taken together, our entire GCN framework is given by:

$$\text{Energy Prediction} = \text{GMP}(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}(\text{ReLu}(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(0)}W^{(0)}))W^{(1)}). \tag{1.3}$$

1.4.1.2 GCN GB Energy Prediction Graph Structure Findings. We then applied this GCN model to our dataset for energy prediction. Our goal with predicting GB energy was twofold, we wanted to tune our model for optimal predictions while also being able to take away lessons about the underlying graph structure of the GB.

GCN Performance Under Node Removal. We next tested our GCN model for prediction effects from removing atoms from the GB graph structure. We used the centro-symmetry parameter to do this. Atoms with lower centro-symmetry values are more likely to be part of a FCC structure and thus be farther away from the center of the GB. This greater distance may indicate those atoms are less impactful in the GB energy value. We removed atoms from the graphs by removing a percentage of the atoms with the lowest centro-symmetry parameters in the graph, thus if we chose a cutoff of 50% of the atoms, we removed the lowest half atoms from every graph according to the centro-symmetry parameter. This made the centro-symmetry cutoff different for every graph but ensured we worked with a consistent percentage of atoms for each graph. We did this for cutoff values from 1 to 98, and plotted the results in Figure 1.3.

From Figure 1.3, we see that there is an optimal percentage of the GB to include in the prediction for our GCN model. Removing nearly 80% of the nodes in each GB resulted in an improvement in our R^2 of more than .15, an MAE improvement of over 15%, and an improvement in MSE of over 30%. Thus in the case of our GCN model, less data may be better for predicting GB energy values. It is important to note that these cutoff percentages are relevant for this dataset only, other datasets may include differing amounts of atoms as a ratio to the length of the GB. However, the takeaway that the GCN has a performance drop if we include too many atoms is general. Plots for MSE and MAE are given in Figure 1.12.

We also tested node removal following this process of removing nodes according to their

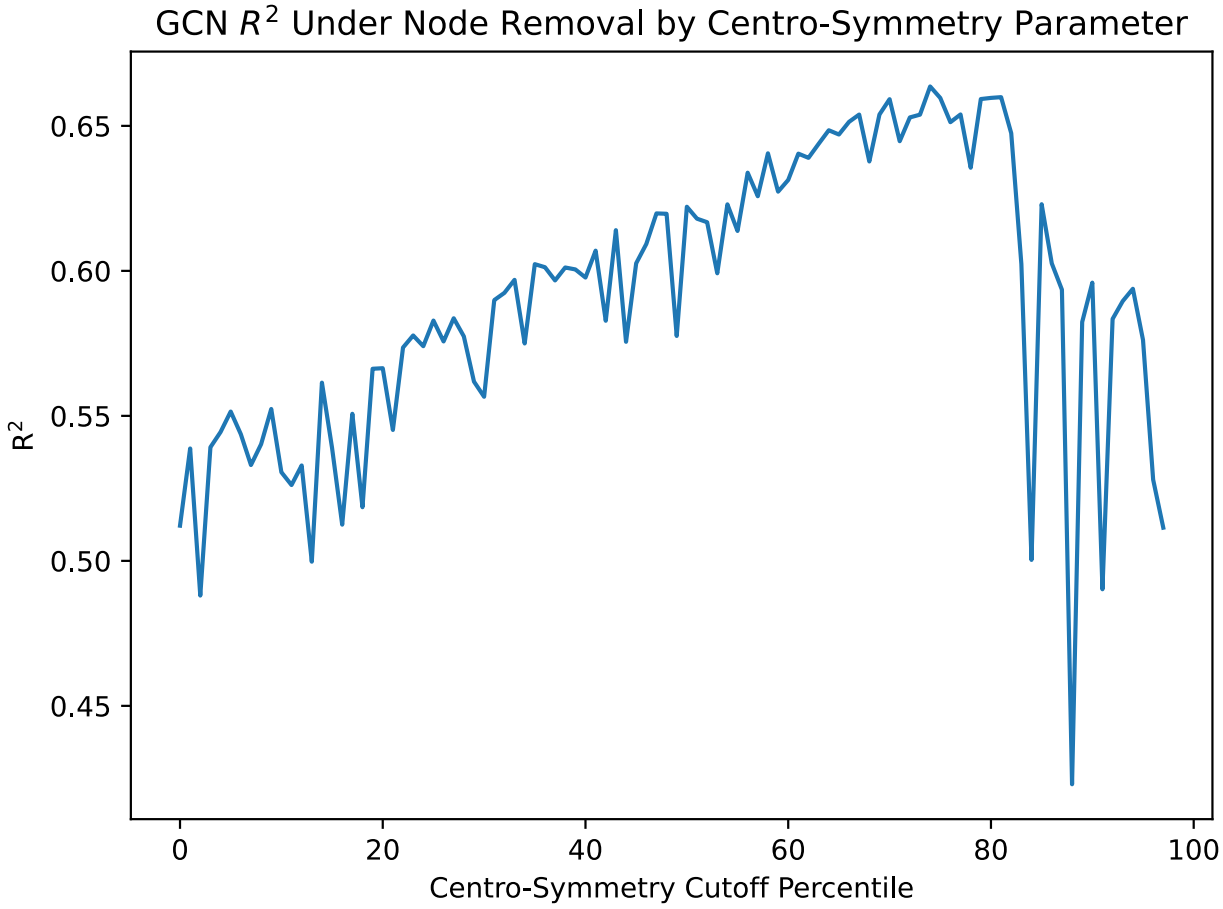

Figure 1.3: GCN performance under node removal. The centro-symmetry cutoff percentile represents the percentage of the atoms we removed from the GB graph according to the lowest centro-symmetry values. We did this for percentiles 1 to 98. This resulted in roughly linear performance increase up until we removed around 80% of nodes, at which point performance volatility increased. However, even when only using 2% of the data, the model outperformed predicting the mean GB energy value of the dataset. Similar plots for MAE and MSE are in Figure 1.13. These plots were derived from graphs created from an epsilon of 3.5 Ångstrom.

centro-symmetry parameter on various epsilon values in the graph creation process. These results are seen in Figure 1.10. We tested removing none, 25%, 50%, and 75% of the data with epsilon values ranging from 2.9 Ångstrom to 4.0 Ångstrom, and performance remained consistent for each cutoff percentage. Also, every cutoff percentage increase provided a marked increase in model performance. Thus our findings in Figure 1.3 are likely applicable to a range of epsilon values in the graph creation, and not just an epsilon of 3.5 Ångstrom as was used to create the figure.

GCN Predicted Energy Values vs True Values. Here we provide a plot of the GCN model's predicted GB energy values vs the true GB energy values, Figure 1.4.

In Figure 1.4 we see visually the sparsity of the lower quartile of GB energy values discussed in Section 1.3, and the density of the values around the mean. In the GCN Optimal Epsilon in Ångstrom for Graph Creation Section, we identified that our model outperformed predicting the mean, a potential pitfall for a dataset like this. Figure 1.4 allows for a visual understanding of this concept, with the model predicting lower GB energy values well with a slight error for generally predicting higher energy values on the low end. Also, the GCN was able to distinguish true GB energy values near the mean, demonstrating the model's performance across all types of GBs included in the dataset.

1.4.1.3 GCN Node Saliency Map. After we verified the performance of our model, we moved onto the task of using our model to distinguish individual atom importance within the GBs. We did this by creating saliency maps of the GBs. Saliency maps work by tracing the gradients of the GCN model by applying backpropagation to the test dataset, and assign a value to each node representing its effect on the output. Higher values indicate a larger effect. This method calculates saliency scores for all atoms in our test dataset, hence we have been able to create importance rankings for GB energy at the level of the atoms. We include saliency maps for multiple GBs in Figure 1.5 and Figure 1.6.

From Figure 1.5 and Figure 1.6, we can visually infer the important areas of each GB displayed. Noteworthy is the variety of areas in the GBs lightened up by the saliency

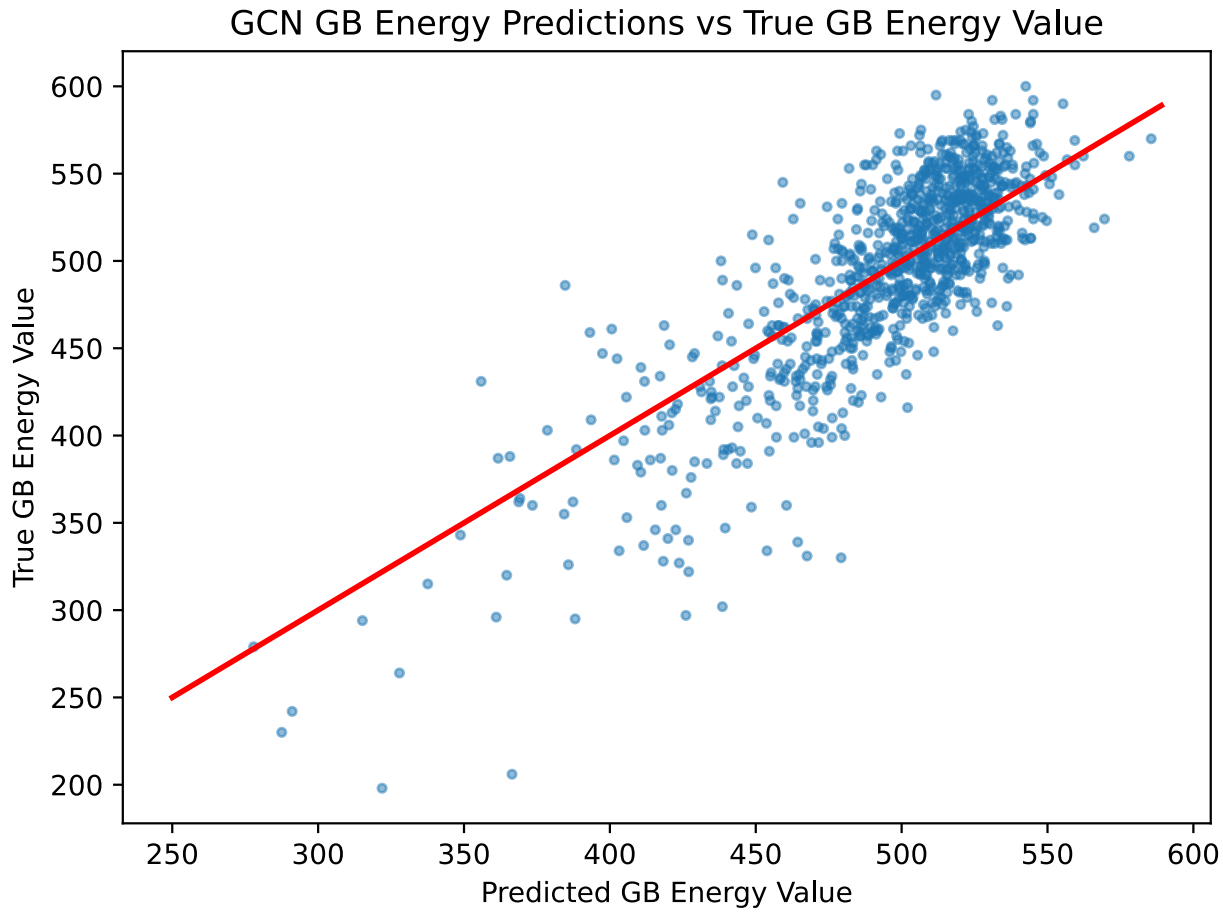

Figure 1.4: GCN predicted GB energy values vs true GB energy values. Our GCN model was capable of learning GB graph qualities affecting GB energy, as seen in Figure 1.3. Here we see this performance was fairly consistent across all GB energy values in the dataset. This GCN model was trained with 75% of the atoms removed according to their centro-symmetry parameter as explained in the GCN Performance Under Node Removal Section.

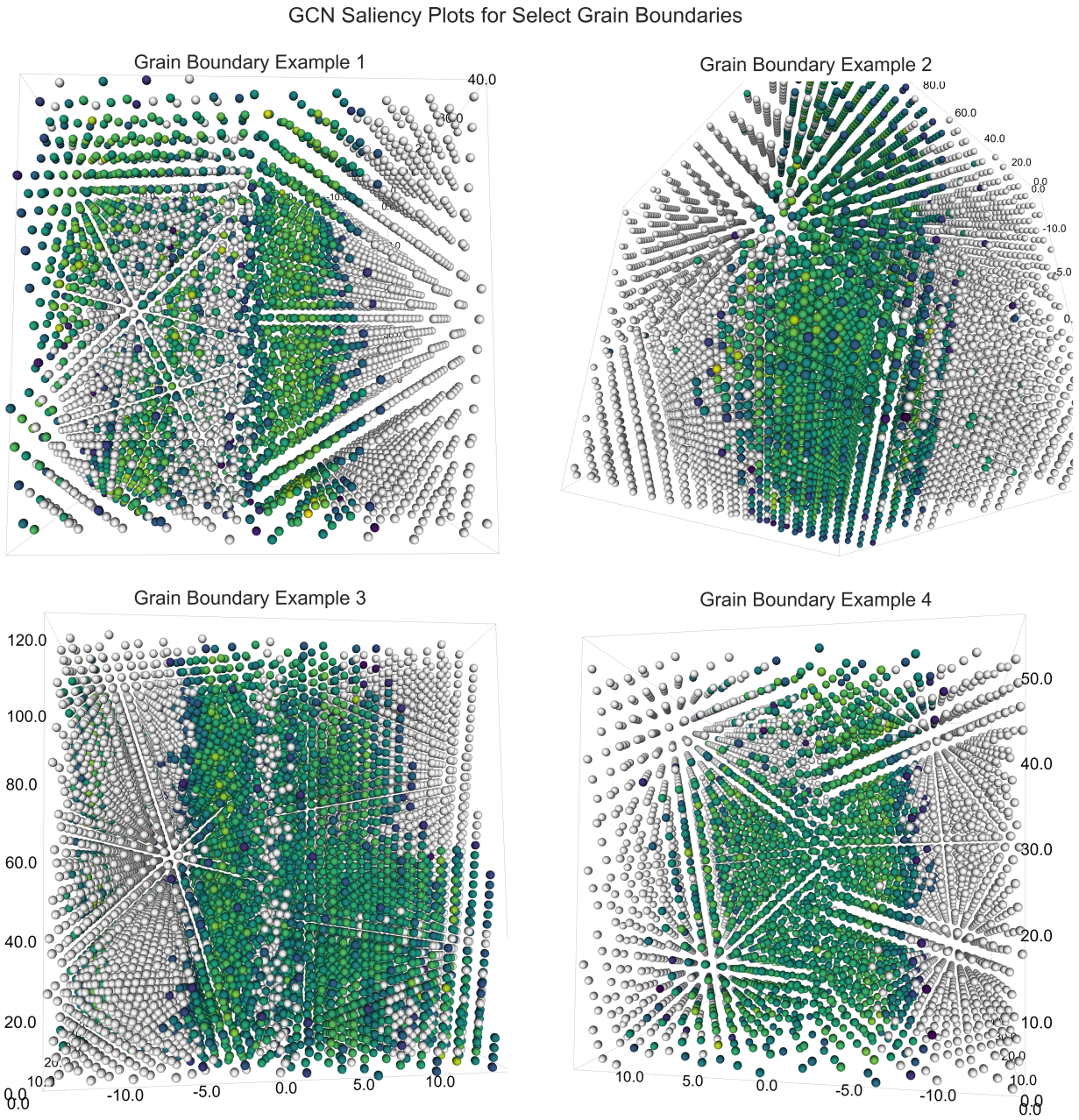

Figure 1.5: GCN saliency maps for select GBs. Lighter colored nodes represent higher saliency. Gray nodes were removed from the learning process according to centro-symmetry parameter values as explained in the GCN Performance Under Node Removal Section with a 50% cutoff rate. Further discussion of the saliency maps can be found in the caption of Figure 1.6.

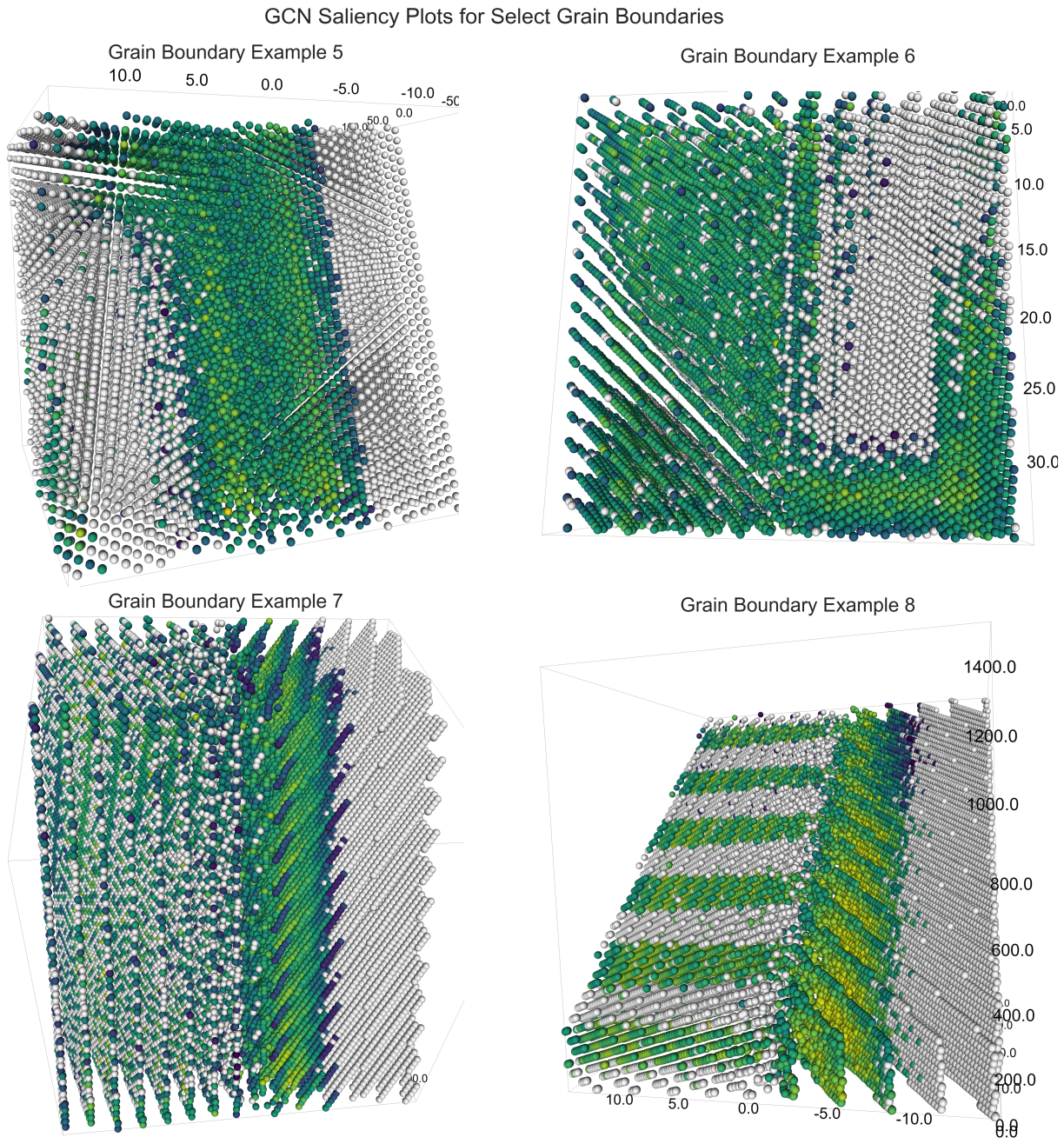

Figure 1.6: GCN saliency maps for select GBs. Lighter colored nodes represent higher saliency. Gray nodes were removed from the learning process according to centro-symmetry parameter values as explained in the GCN Performance Under Node Removal Section with a 50% cutoff rate. We chose to display these saliency maps due to them being representative of common saliency maps inside the test dataset. GBs such as examples 1, 3, 5, 6, 7, and 8 have at least one crystallite that displays high atomic importance for atoms near but not directly along the center of the GB. Another observation is while some crystallites had high centro-symmetry parameters near the GB as expected, others have a more even spread of centro-symmetry parameters throughout the crystallite such as the left crystallite in Grain Boundary Example 7.

maps; some crystallites cluster importance near the center of the GB but not directly along the boundary, while others have a more even spread across all atoms used in the model. While this visual representation is useful, further research into the atoms that produced high saliency outputs would add more to the discussion of which atoms are likely to influence GB energy and why that is the case.

1.4.2 Graph2vec. The graph2vec algorithm was first introduced in 2017 by Narayanan et al. This algorithm builds on the doc2vec algorithm, which in turn builds off of the word2vec algorithm [31, 28]. Graph2vec has considerable strengths, foremost of these is that the embeddings are unsupervised, meaning the embeddings we created with the algorithm may apply well to other classification or regression tasks without the need to retrain the model. The graph2vec algorithm is based solely off of local node geometries, implying all learning we find in our embeddings is based solely off of the graph structure itself. Since graph2vec does not use atomic coordinates as a feature, this method is rotation and permutation invariant, satisfying the necessary requirements for a GB embedding.

1.4.2.1 How Our Graph2vec Algorithm Works. The graph2vec algorithm begins by representing each GB graph as a document. This is done by creating a Weisfeiler-Lehman (WL) hash for a rooted subgraph based at each node [40]. Hence these subgraphs, one based at each node, create the vocabulary for our graph, or document. The WL hash method takes into account a node feature, in our case node degree, and a neighborhood distance to use from the base node to create the hashing. From our testing found a maximum distance of two from the base node was optimal, which involves performing two iterations of the algorithm per node. The WL hashing algorithm then starts at the base node and subsequently aggregates the base node's hash with all other nodes inside the permitted neighborhood. This creates a hash for every rooted subgraph with our base node as the root. The WL hashing algorithm guarantees isomorphic subgraphs will have the same hash and nearly guarantees that non-isomorphic graphs will not. Research has shown that the rate that the WL hashing algorithm gives the same hash to non-isomorphic graphs is low enough to not affect the majority of

datasets, especially when at least two iterations are performed [49]. Therefore our document representing our graph contains a vocabulary uniquely identifying every subgraph structure within the GB. Thus for each GB we have created a document that captures the underlying graph structure.

The next step in the algorithm is to train the neural network and update the embeddings. Every graph is randomly initialized in our embedding space, in our case 128 dimensions. The training then follows the skip-gram model and negative sampling. We will now describe this process and give the general steps we follow for every training epoch. Each epoch begins by shuffling all GB graphs. After that, the algorithm passes through each graph, which we will call the current graph, and updates its embedding. The neural network is then updated through the Skip-gram method.

The Skip-gram method begins by selecting a random subset of words, or subgraph hashes, from the current graph's document, and then training the neural network on predicting the probability of each word based on the current graph's embedding. We are trying to predict the context of the graph, or the graph's subgraphs, from its embedding. The neural network consists of one hidden layer with a set of weights leading into a softmax function. The graph2vec then uses stochastic gradient descent to update the weights of the neural network for each of these subgraph current graph pairs. Once this training process is complete, graph2vec uses negative sampling to update other graphs' embeddings. This process is done by selecting several subgraph hashes that are not contained within the current graph's document. For each of these negative samples, a graph's embedding in which they reside is passed into the neural network with the subgraph hash to get a probability of the subgraph hash being in the graph's document. The embedding of the graph is then updated through stochastic gradient descent in order to maximize the probability of the embedding successfully predicting the subgraph hash. This is done for every negative sample, then the next graph in the list is made the current graph, and the process repeats itself. The goal of graph2vec embeddings is to have similar graphs close together and different graphs far apart. It follows

that the dimensions in our embedding represent GB characteristics.

1.4.2.2 Graph2vec GB Energy Prediction Graph Structure Findings. We now discuss our results and insights gained from applying graph2vec to our GB dataset. We used the graph2vec implementation found in the Karate Club package for all tests [39]. Since graph2vec produces fixed length embeddings of the GBs, we used ridge regression on the embeddings themselves to compute all statistics in this section.

Graph2vec Performance Under Node Removal. We next tested our graph2vec algorithm performance under node removal. We followed the same process to remove nodes here as we did in the GCN Performance Under Node Removal Section. We used the centro-symmetry parameter, an indicator of whether a node is FCC, or likely to be away from the center of the GB, to remove nodes. We tested graph cutoffs for 1 to 98 percent of the data, the percentage indicating the percent of atoms that were removed from each graph due to having lowest centro-symmetry values. The results are in Figure 1.8.

From Figure 1.7 we see how weak our graph2vec algorithm is to node removal. Performance in every indicator fell, even at low removal rates. Once we had removed 20% of the data, the graph2vec model was not learning anything regarding GB energy values. This likely has to do with graph2vec's negative sampling method for learning graph embeddings. Since graph2vec only updates graphs' embeddings according to a negative sample from a base graph, using all the atoms in GBs should have many FCC structures. FCC structures generally appear away from the GB, hence have low impact on the GB energy value and may be best left out of the training process. Once we began removing atoms, we likely began removing atoms from these FCC structures within the GBs and began creating subgraph that either appeared unique or like subgraphs that appeared along the GB. This would have made it difficult for the model to identify subgraphs that were influencing the GB the most. This negative sampling technique may be a strength of the graph2vec algorithm when presented with the complete data; it is probable the algorithm avoids using FCC subgraphs in the embedding process. Plots of MAE and MSE performance are given in Figure 1.13.

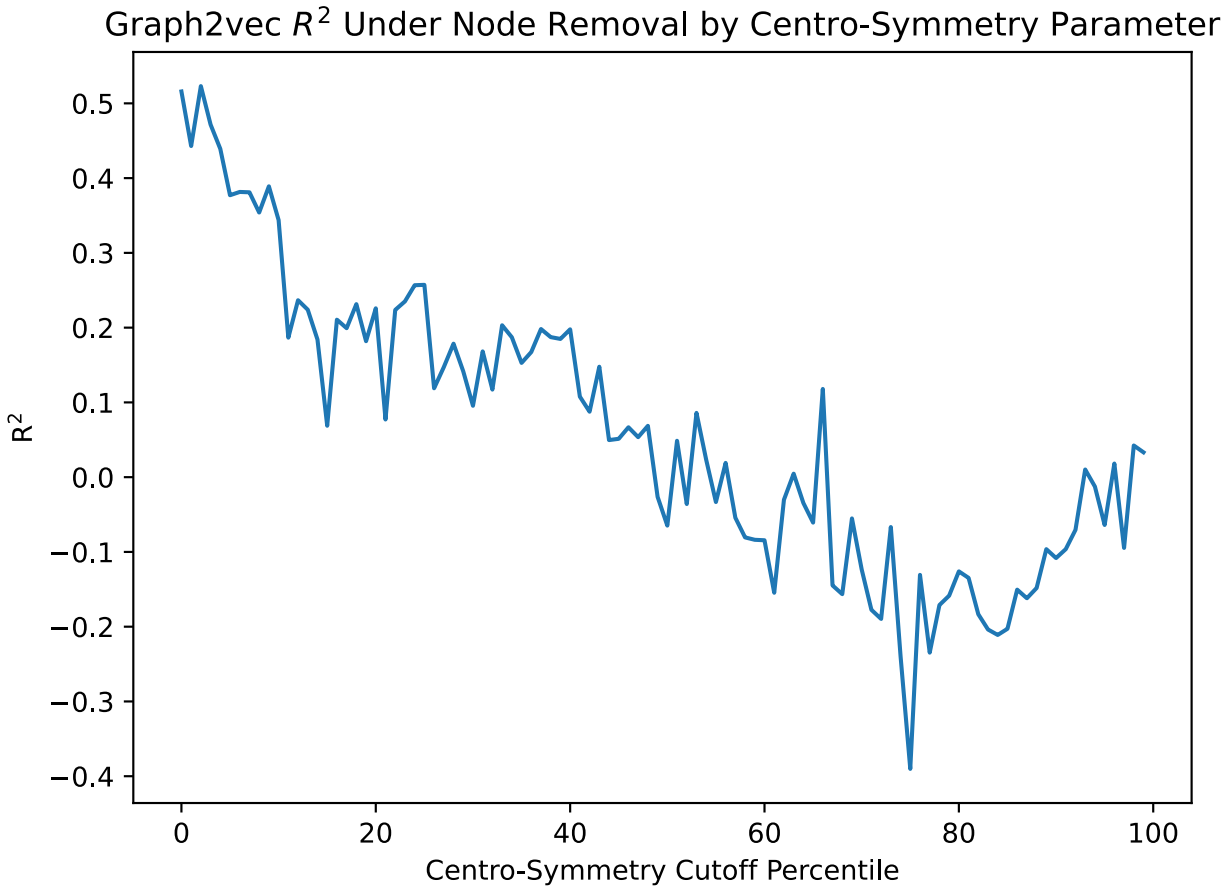

Figure 1.7: Graph2vec R^2 performance under node removal. Atoms were removed from the graphs according to their centro-symmetry parameter, with the atoms with the lowest values being removed. Centro-symmetry may be an indicator of being near the center of the GB. Performance began to drop almost immediately, with noticeable drops in performance happening within 5% of the atoms being removed. Once 20% of the data was removed, performance indicates the model was not able to learn. Similar plots for MAE and MSE are in Figure 1.13. These plots were calculated on graphs created from an epsilon of 3.5 Ångstrom.

Graph2vec Predicted Energy Values vs True Values. We now discuss graph2vec GB energy value predictions created from applying ridge regression to the graph embeddings vs the true GB energy values. This is shown in Figure 1.8.

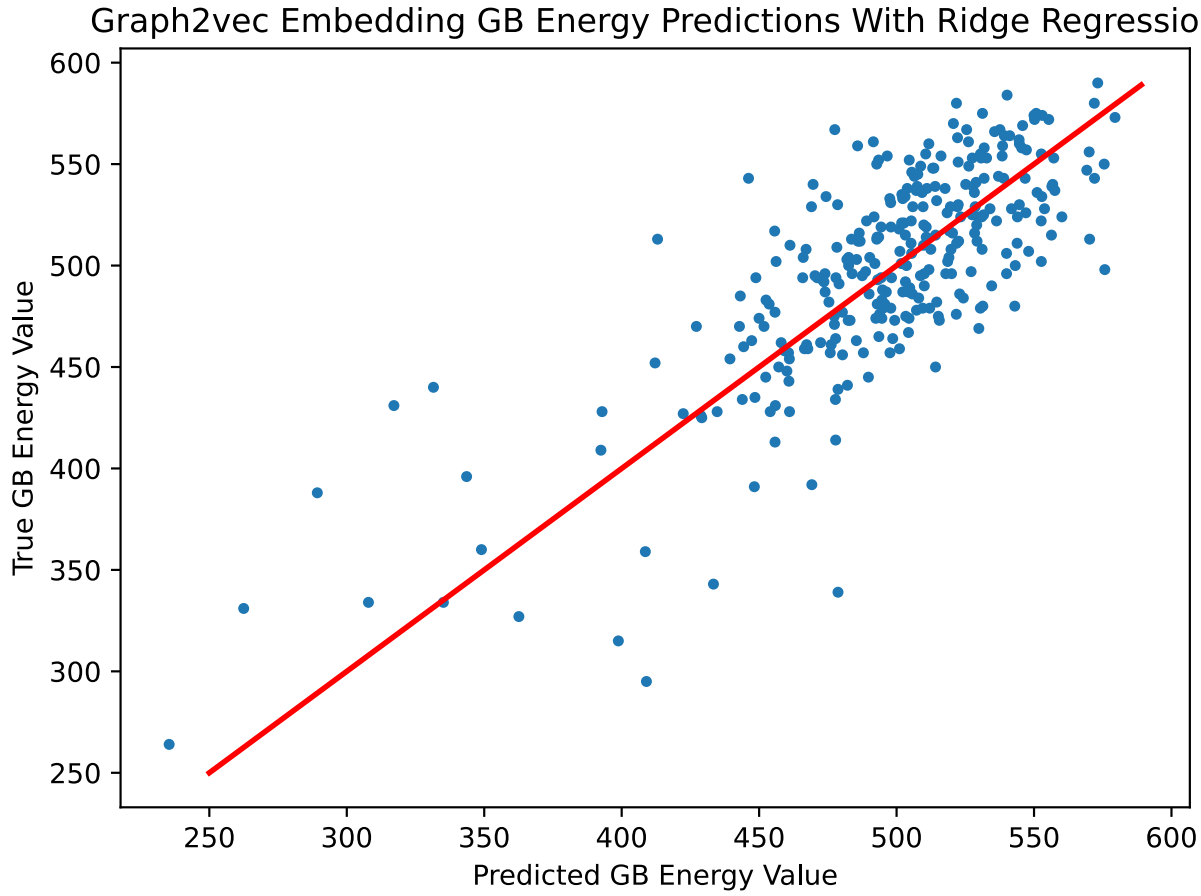

Figure 1.8: Ridge regression applied to graph2vec GB embeddings in order to predict true GB energy values. Ridge regression was able to properly identify differences encoded in the embedding indicating GB energy value, seen by the algorithm correctly separating several of the low energy grain boundaries from the main mass of energy values near the mean GB energy. Applying ridge regression to the graph2vec embeddings also performed well in the main mass of values near the mean, separating several GBs well that had similar energy values.

Figure 1.8 visualizes our results from Figure 1.11, which gave strong evidence to our graph2vec encoding GB energy value information in the graph embedding. Here we see how successful ridge regression applied to our graph2vec embeddings was at separating low energy GBs from higher energy GBs, those with energy values less than 472 and all GBs with

energy values above 472 respectively. We also see how successful the ridge regression was at spreading out GBs that had true energy values near the mean. Since graph2vec only learns from negative sampling applied to graph subgraphs pairs based on graph structure alone, this result gives evidence that graph2vec was able to identify certain subgraph structures that were more common in the low energy GBs from the small sample size in this category. The even spread through higher energy GBs also indicates graph2vec was able to identify slight differences in subgraph structure that led to these slight changes in energy values within the more densely clustered energy values. Thus graph2vec not only captured GB energy values within its unsupervised embeddings, graph2vec also located the important subgraphs within these graphs regarding GB energy values.

1.4.2.3 Graph2vec Subgraph Importance Ranking. Now that we established that graph2vec successfully captures GB knowledge in the embeddings and uses important subgraphs regarding GB energy in the training process, we focus on our goal of identifying these important subgraphs. To do this, we added a method from the Doc2vec base class to our graph2vec algorithm that allows subgraphs to be embedded in the same embedding space as the overall graphs. This method does not embed all subgraphs, only the ones that are used in the negative sampling training process. Since the goal of graph2vec embeddings is to have similar graphs embedded near each other, subgraphs whose embeddings appear close to graph embeddings in which they reside could be indicative of the overall graph structure and could have been the most influential in the embedding process. We measured similarity between subgraph embeddings and graph embeddings through cosine similarity. We display some similar subgraph GB pairs in Figure 1.9.

While many GB embeddings did not end up close to any of their subgraph embeddings, in the case of the grain boundaries in Figure 1.9 each GB embedding had a cosine similarity score of at least .6 with the top 10 subgraphs ranked by cosine similarity. For both examples, we see that the subgraphs whose embeddings had the highest cosine similarity to the overall graph embedding tend to cluster near the edges and form chains throughout the graph. In

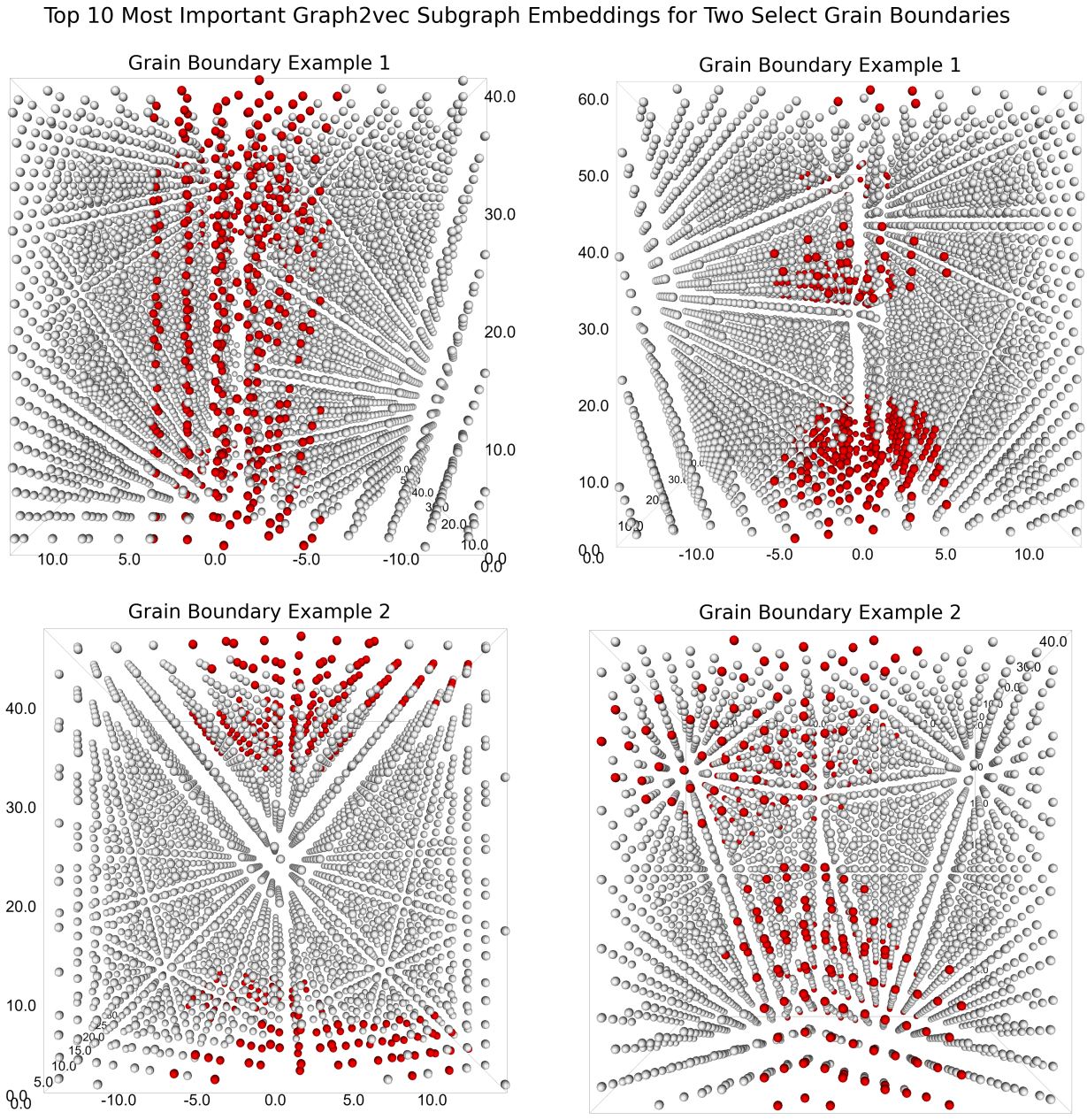

Figure 1.9: The top 10 closest subgraph embeddings for two select grain boundaries from the test set. In each row the image on the right is rotated 90 degrees down from the image on the left. For GB Example 1, the subgraph embeddings had a cosine similarity score of .63-.65 compared to the graph embedding. In GB Example 2 the embeddings had a cosine similarity score of .6-.75. Notice how the subgraphs formed a chain in Example 1 and how the subgraphs were clustered at the edges in each example.

each case the majority of the subgraphs had nodes from each metallic crystallite in the GB. Since graph2vec only chooses negative samples during the embedding training process, the subgraphs seen in the figure may be less common in other grain boundaries, and may be capturing the areas of the GB that are uncommon across the dataset.

A reason that many graphs did not have any subgraph embeddings with a high cosine similarity score may be due to each subgraph capturing a different aspect of the GB structure. Taken as a group, these subgraph embeddings might have captured the most important aspects of the GB in which they reside and explain the overall structure well, with each subgraph being near to the GB in the embedding space for a few dimensions. Further research into this area would be necessary in order to establish whether or not this is the case and how these subgraph embeddings relate to the graph embeddings.

1.5 Conclusion

Our GCN and graph2vec models were effective at reaching our goals of predicting GB energy values and learning information about the atom level structure. Each model presented unique strengths in applications of reaching these research goals, providing diverse insights on the GB structure at the level of the atoms.

Our GCN model was not only adept at handling atom removal from the GB, its performance improved when we removed up to 80% of the atoms with low centro-symmetry parameters. In contrast, the graph2vec model performed as poorly as predicting the GB energy mean after only 20% of the atoms with low centro-symmetry values were removed from each graph. The GCN model had better scores than graph2vec when comparing R^2, MAE, and MSE for GB energy value predictions, but large improvements over graph2vec were only seen with atom removal. If we included all atoms in the GB graph, the GCN model was only slightly better than graph2vec, according to these performance statistics.

GCN was particularly useful for understanding the GBs from their atoms. We created saliency maps through backpropagation that allowed us to rank nodes based on their in-

fluence in predicting the GB energy value. The saliency values for each atom were key to gaining a better understanding of the GBs at this atom level, one of our main research goals, and open up a range of future research opportunities regarding their properties and placement in the GB.

Graph2vec was a unique take on GB embedding methods, as graph2vec turned a GB into a document before embedding. A strength of graph2vec is that the model was unsupervised as opposed to semi-supervised for our GCN. Since graph2vec was unsupervised, the embeddings the model created could contain further graph knowledge than just GB energy information and could be useful for a variety of applications. Another possible strength for graph2vec in comparison to our GCN model was graph2vec solely learned based on graph structure, where the GCN model took in five parameter values for each atom to perform the convolutions; however, three of these five parameters were required to create the graphs to learn on.

Since graph2vec learned from graph structure alone and has the ability to embed subgraphs in the same space as the GB graphs themselves, the model might have been able to identify key subgraphs from each GB in its learning process. This ability allows for a deeper understanding of which subgraphs are influential in the GB. From our further testing we know this influence at least extends to the GB energy value. A deeper look into these subgraphs would allow us to gain a better understanding which areas of the GB influence energy values or other properties from the atoms themselves, thus helping us achieve our goal of understanding GBs from the level of their atoms. This understanding is in addition to what we learn from the GCN saliency scores, since saliency scores rank individual atoms by importance and graph2vec identifies likely important subgraphs within the GB.

1.6 Appendix

1.6.1 Finding Optimal Epsilon for Graph Creation. Here we provide in-depth on our decision to use epsilon equal to 3.5 Ångstrom for our graph creation. In our tests of several epsilon values, we did not test values less than 2.9 due to problems with some GB graphs being completely disconnected. We also did not test values greater than 4.0 due to computational time required to create the graphs.

1.6.1.1 GCN Optimal Epsilon in Ångstrom for Graph Creation. We began by testing for the optimal epsilon in Ångstrom for our graph creation. We did a search through several epsilon values for our epsilon nearest neighbors graph creation applied to the same train, test, and validation GBs. Hence the only difference in our epsilon search was due to the change in epsilon. These results are in Figure 1.10.

To compile the results for Figure 1.10, we included a node removal for the centro-symmetry parameter. This was added after our initial tests for the optimal epsilon due to our findings in the GCN Performance Under Node Removal Section and will be discussed there. Here we focus on the results where no data was cut from our graphs during the epsilon search process, which is the line labeled no data cut in the figure. We tested epsilon values from 2.9 to 4.0 Ångstrom, which resulted in a mean degree increase from 9.3 to 11.4 across the dataset. Our tests resulted in consistent performance across all epsilon values, remaining around .52 for our R^2 score, 31 for the mean absolute error (MAE), and $1,600$ for the mean squared error (MSE). This is far off from the results the authors of the dataset received from applying SOAP to the dataset, but our results are sufficient to prove our GCN model has been able to learn features about GBs. If our model were to predict the mean GB energy values, it would have an MAE over 41 and a MSE of more than $3,000$, twice our model's MSE.

Thus we see that our GCN model is robust at GB energy predictions for a large range of epsilon values in our graph creation. This robustness gives us reason to believe the GCN model is resilient to small perturbations in the data as well. If an atom were to move slightly

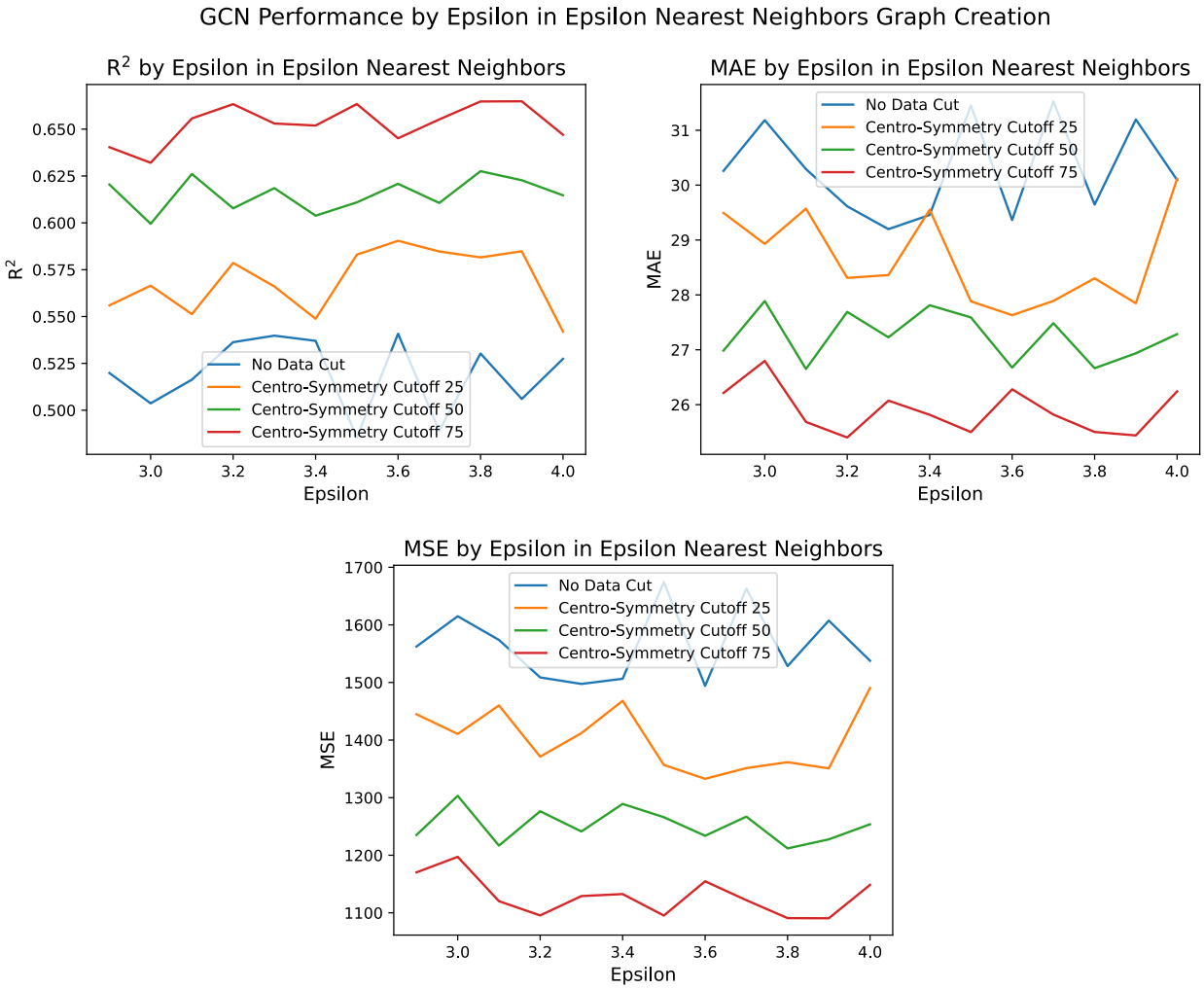

Figure 1.10: Here we tested several values for epsilon in our creation of GB graphs through epsilon nearest neighbors. The GCN is resilient to changes in the underlying graph structure, with the R^2, mean absolute error (MAE), and mean squared error (MSE) remaining relatively consistent across the board from an epsilon of 2.9 to 4.0 Ångstrom. We also show how removing atoms by their centro-symmetry parameter helps our model learn the important areas of the graph structure for all epsilon values, a finding discussed in the GCN Performance Under Node Removal Section. For each line, the cutoff number refers to what lower percentile of atoms were cut out of the graph according to their centro-symmetry parameter. Generally atoms in an FCC position have lower centro-symmetry values. Remarkably, cutting out 75% of the atoms according to this method increased R^2 by roughly .15, as well as MAE by 16% and MSE by 30%.

out of place and fall out of position for several edges, or move into position to add several edges, such differences likely would not affect our model's performance.

1.6.1.2 Graph2vec Optimal Epsilon in Ångstrom for Graph Creation. To begin testing our graph2vec model, we first did an epsilon search to find the optimal epsilon for our graph2vec model to use for our epsilon nearest neighbors graph creation. We measured performance with R^2, MAE, and MSE to find the optimal epsilon value. We tested epsilon values from 2.9 to 4.0 Ångstrom. All tests were done of the same datasets, the only difference being the epsilon value. The results are in Figure 1.11.

In Figure 1.11 we see how impactful the training process is for the graph2vec and highlight several weaknessness of the algorithm. If a subgraph has not been visited in the model training, it cannot be used in the embedding of a test graph. This is due to graph2vec viewing subgraphs as words; if it has never seen a word before it has no understanding of the meaning. However, if the model is applied to several very similar grain boundaries, this may turn into a strength. During the training process graphs are trained through negative sampling, so this model may perform well at picking up on subtle GB differences. In this case we used the same training and testing data that we used for the GCN model tests, but other test, train, validation divisions in the data often struggled to perform with an epsilon above 3.9 Ångstrom. That said, model variability for graph2vec likely has more to do with the training process and the data used for training and testing than it does with the epsilon value chosen for epsilon nearest neighbors graph creation.

Since epsilon equal to 3.5 Ångstrom resulted in stable performance for both models we chose to use it for our model comparisons.

1.6.2 MAE and MSE During Node Removal.

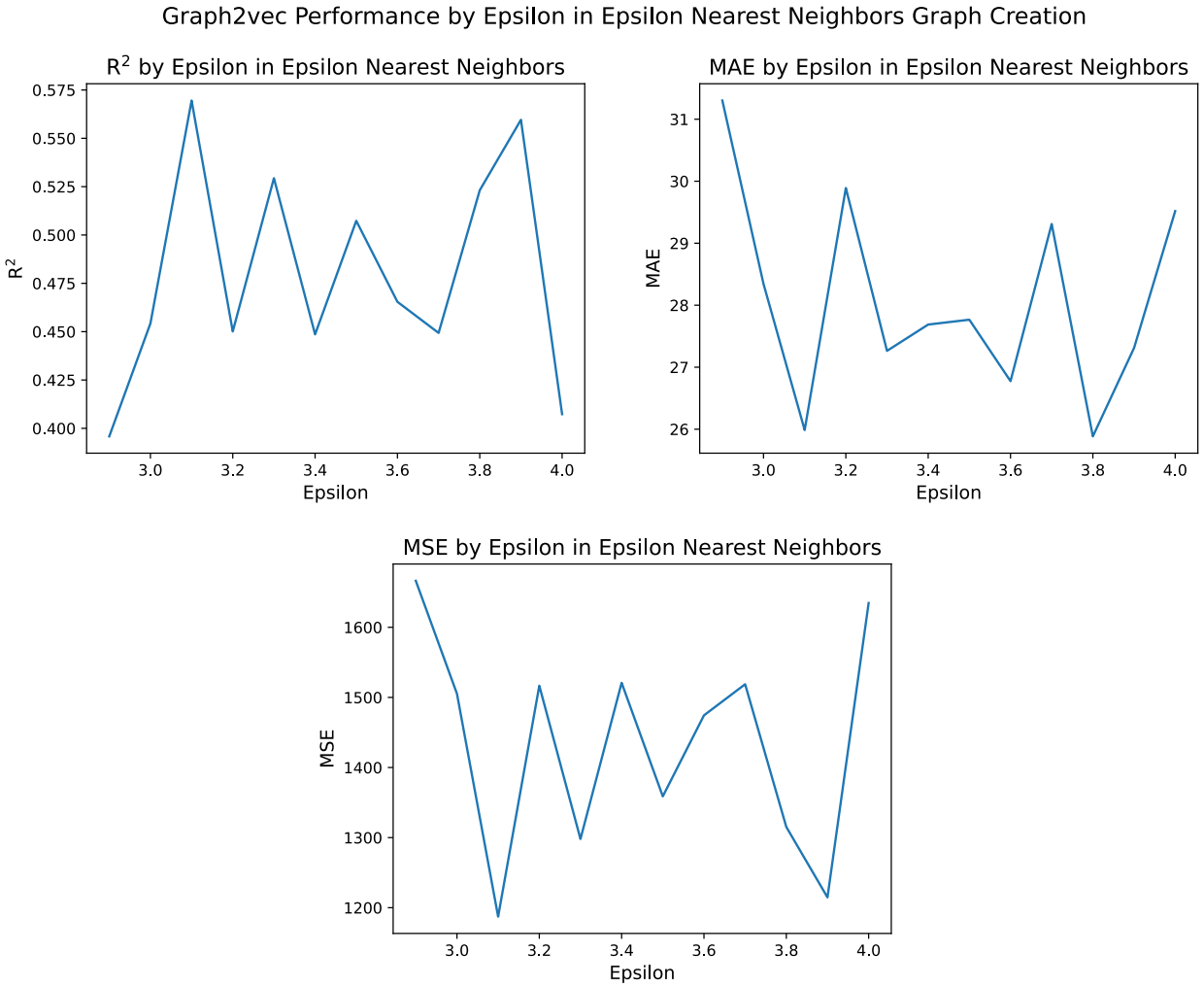

Figure 1.11: Graph2vec performance by epsilon in epsilon nearest neighbors graph creation for epsilon 2.9 to 4.0 Ångstrom. All values reflected high variance in the results around their means. This highlights a weakness of the graph2vec algorithm. The training process does not view all subgraphs included in the training data, and since the subgraphs are viewed as words, if a subgraph in the test data does not appear in the training data, graph2vec cannot use that subgraph in the classification process.

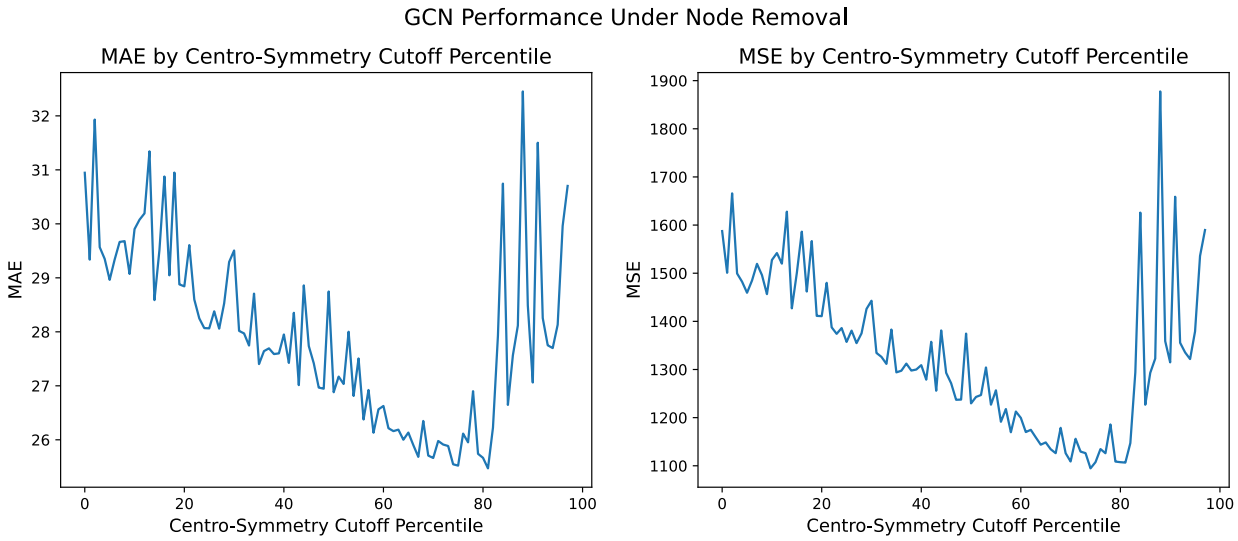

Figure 1.12: GCN MAE and MSE performance under node removal. The centro-symmetry cutoff percentile represents the percentage of the atoms we removed from the GB graph according to the lowest centro-symmetry values. We did this for percentiles 1 to 98. This resulted in roughly linear performance increase up until we removed around 80% of nodes, at which point performance volatility increased. However, even when only using 2% of the data, the model outperformed predicting the mean GB energy value of the dataset, which would result in a MAE of 41 and a MSE of more than $3,000$. These plots were derived from graphs created from an epsilon of 3.5 Ångstrom.

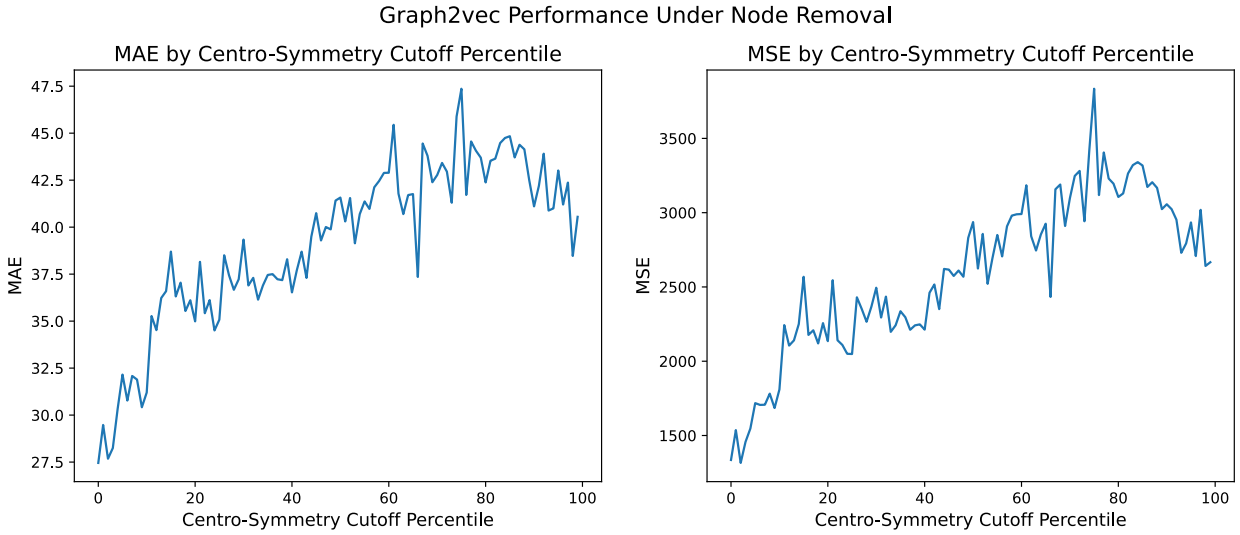

Figure 1.13: Graph2vec MAE and MSE performance under node removal. Atoms were removed from the graphs according to centro-symmetry parameter, with the atoms with the lowest values being removed. Centro-symmetry may be an indicator of being near the center of the GB. Performance began to drop almost immediately, with noticeable drops in performance happening within 5% of the atoms being removed. Once 20% of the data was removed performance was at or below expected performance from predicting the mean, which would result in a MAE of 41 and a MSE of more than $3,000$. These plots were calculated on graphs created from an epsilon of 3.5 Ångstrom.

Chapter 2. Global Supply Chain and Strategic Alliance Networks

2.1 Introduction

Businesses around the world provide people with services and products they need in order to go about their daily lives. There is hardly a service or product offered that is not built by a network of several businesses working together, such as suppliers, creditors, and landlords. Even just one company failing to produce their product or service can have outsized effects on the overall economy, even affecting those who do not use their product or service [12]. Expanding this network of companies creating products and services together to all companies in the world constitutes the world's supply chain network. As part of the global supply chain, many companies produce products and services jointly, each performing an essential part of the project while remaining independent companies. These relationships are known as strategic alliances, and the combined global supply chain network and global strategic alliance network is called the global value chain network.

Disruptions to the global value chain network are inevitable, and businesses and governments alike need to proactively prepare for these inevitable value chain disruptions. By thoroughly understanding the global value chain dynamics and structure, businesses can curb losses and governments defend against economic weaknesses and threats [24]. In this chapter, we conduct a rigorous mathematical examination of the structure and attributes of the dynamical global value chain to address this need for better understanding and to identify the dynamic value chain's behavior. However, the necessary structural and dynamic analysis cannot be performed without proper data, and global value chain data is thoroughly lacking, especially on the academic stage.

We present the following dataset to fill this void. To monitor the current global situation, watch for potential disruptions, and identify trends that can upend the economy, we compiled a dynamic global supply chain dataset with detailed business information and relationship

data, along with a global strategic alliance dataset containing detailed business information. These two datasets construct the global value chain with over $225,000$ companies. This global value chain dataset is updated almost daily. The frequent updates enable analysts to rapidly identify trends, such as supply chain reshoring, which is when businesses move their supply chain to their home country from a foreign country, shocks propagating through the network, and other valuable insights into the global economy previously hidden from view [38].

We contribute the following in this chapter:

- A dynamic global value chain network unique collectively in scope and accessibility.
- Up-to-date applicable business findings made possible only through our dynamic global value chain dataset.
- Global value chain community structure results and insights from our global value chain dataset.

2.2 Background

While supply chains are highly researched, in our literature review we did not come across anything similar to our dataset regarding its scope and dynamic nature. Due to significant differences between academic and commercial datasets we cover these areas separately.

2.2.1 Commercial Supply Chain Datasets. There are three significant commercial dynamic supply chain datasets. They are provided by Panjiva, Bloomberg, and FactSet. Panjiva's dataset, which is one of the largest available, pulls from company shipping records and is available for a fee. The dataset contains approximately $13,000,000$ company-to-company relationships with trade volumes from customs paperwork [15]. Our dataset has fewer listed relationships than Panjiva's and involves fewer companies. However, ours brings significant strengths in areas where Panjiva's lacks. We include relationships that do not

involve shipping transactions, such as landlords, creditors, and domestic relationships. Additionally, the Panjiva supply chain dataset is limited to shipment records from 17 countries while our dataset is global in scope. Our dataset also contains global strategic alliance data, which Panjiva does not have at all. Thus our dataset is more comprehensive than Panjiva's dataset and allows for more comprehensive research of the global supply chain situation.

Bloomberg provides a global supply chain dataset for a fee, upon inquiry, we were quoted at $\$100,000$ annually. Bloomberg has historical supply chain data on $123,000$ companies going back to 2006 with $450,000$ relationships. This dataset has more historical data than ours, but contains fewer companies and does not contain any strategic alliance data. Bloomberg provides relationship value estimates for $1,300$ companies, while we have relationship labels for all company relationships. Another key difference is Bloomberg's dataset is geared towards helping companies analyze changes to their own supply chain; from our research we found no way for a client to download the entire dataset as a dynamic network. It appears that Bloomberg gives clients access to the specific data they are interested in, not the entire dataset. Our dataset is readable in Python and is already contained in a dynamic format. There is no need to request specific parts of the data through customer service as there is with Bloomberg, providing faster and simpler accessibility [6].

The third and final commercial dataset that tracks the global supply chain is the FactSet supply chain dataset, also provided for a fee. This dataset stretches back to 2003 for North America and subsequently includes other areas of the world until including all by 2016. However, the dataset is still heavily focused on North America, and is not as globally rounded as our dataset. The FactSet dataset includes roughly $31,000$ companies, which makes it roughly one-sixth the size of our dataset. Factset tracks company relationships by listing other companies as either a customer, supplier, partner, or competitor. Their relationship data differs by directly labeling how companies interact with each other through the supply chain even indirectly, while our dataset labels the direct business relationships between firms, and indirect relationships can only be inferred through analysis [21].

The Panjiva, Bloomberg, and FactSet datasets are the most complete we have found, and all have a cost barrier in order to study them. We created our dataset through S&P CapitalIQ, which we have access to through a fee. Capital IQ is the research branch of S&P Global, which provides ratings, benchmarks, and analytics of the global capital and commodity markets [16]. Capital IQ retrieves their data from public companies financials, private company data, estimates, transactions such as funding rounds and public offerings, credit ratings research, and macroeconomic data [17].

2.2.2 Academic Supply Chain Datasets. Due to the difficulty of accessing and analyzing these datasets, researchers often use other methods to acquire supply chain data. One method is to use value added taxes (VAT) transaction data. Diem et al., 2022, used this method to analyze the Hungarian supply chain network, which contained over $90,000$ companies. A downside to this method is not all countries have VAT taxes, including the United States, thus a large part of the world's economy is excluded with this method [12].

Tokyo Shoko Research provides supply chain data for nearly $1,000,000$ companies in Japan and $3,544,343$ relationships between them and while the dataset is produced commercially, it has been widely used in academic research, such as in [22, 23, 41]. This dataset provides an in-depth view of Japan's supply chain network, but it is limited to Japan as it does not include foreign companies in the supply chain.

Other researchers commonly use either synthetic data or small supply chains, such as Yang et al., 2021, which used both to model robustness of the supply chain in Europe. Their real-world model of a European supply chain contained 38 companies [48].

Overall, within academia the majority of research has either been applied to country-level supply chains, synthetic supply chains, or small supply chains. Our global dataset provides researchers the opportunity to expand their research beyond these constraints.

2.2.3 Supply Chain Research and Studies. Current research regarding supply chain has focused significantly on supply chain robustness, such as how susceptible supply chains

are to shock propagation [48, 24], and resilience, which is how supply chains adapt to adverse events [45].

In the study of supply chain robustness, researchers have focused on economic stability. One study performed on a synthetic European supply chain network measured the importance of maintaining surplus inventory and backup suppliers to avoid the effects of demand shocks [48]. The study involving the Hungarian supply chain network discussed in Section 2.2.2 computed the economic systemic risk of each company in the network and found that company failure impact on the overall supply chain correlated strongly with the company's position within the supply chain. The economic systemic risk can be reduced through supply chain redundancies and changes in network topology [12].

The supply chain network for Japan in Section 2.2.2 has been used to test for both resilience and robustness. One study of this supply chain measured the impact of the 2011 earthquake on the supply chain and used the results to model the effects of other natural disasters. This study found that the largest impact of natural disasters on the supply chain would be from shock propagation, not from the direct effects of the disaster itself [22]. Another study on this same supply chain dataset for Japan measured the impact on the economy if Tokyo were under lockdown. The study found that the effects would propagate quickly through the country and impact the whole country's economy [23]. Other researcher has examined how resilient Japan's supply chain was during a natural disaster, under a lockdown, and at how companies can increase their own supply chain's resilience. They found that geographic diversity and diversity from involvement with other communities in the supply chain significantly increased a company's resilience.

An influential study found that microeconomic shocks, such as shocks to a sector, create sizable aggregate fluctuations across the whole economy only if there are asymmetries in the downstream supply chain of the sector. This research supports our claim that identifying asymmetries within the supply chain network through applications of network theory will identify weak areas of the supply chain network [1].

Our dataset is well-positioned for studying all these areas such as supply chain robustness, resilience, and shock propagation on a global scale due to its scope and frequent updates. In Section 2.4.2, we outline how quickly the global value chain is changing. These changes relate to real-world examples of supply chain resilience and robustness.

2.2.4 Global Strategic Alliance Datasets and Research. We found no evidence of a current or historic global strategic alliance dataset. A majority of strategic alliance research has focused on the benefits of collaboration through strategic alliances [2], dynamic capabilities of strategic alliances [30], and how to choose and manage strategic alliance partners [7, 46]. We can use this knowledge of strategic alliances to analyze the benefits and impacts of the more than $90,000$ strategic alliances contained in our dataset.

2.3 Dataset

This dataset consists of two distinct time series business networks, a supply chain network and a strategic alliance network. We begin with a brief overview of the supply chain network. This is a directed supply chain where edges point from the supplier to the company being supplied. Every supplier relationship has one of eight classifications stored as edge labels: vendor, transfer agent, supplier, creditor, lessor, landlord, franchisor, or licensor. These relationships are not mutually exclusive, for example, a company can be both a vendor and a supplier for another company. This is represented by two separate directed labeled edges. The supply chain network consists of more than $175,000$ firms as nodes and more than $415,000$ relationships.

We now provide a brief overview of the strategic alliance network. A strategic alliance between two companies is a relationship where they "undertake a mutually beneficial project while each retains its independence" [25]. Thus the resulting network consists of deep company relationships where each company in a relationship has inherent motivation for the other company's success in fulfilling their end of the deal. This network is undirected and

consists of more than 65,000 companies and more than 90,000 alliances.

We pull data for these networks almost daily from Capital IQ in order to create our dynamic global value chain [17]. This data is difficult to retrieve from Capital IQ, which makes our dataset exceptional in its scope and completeness. Every quarter we pull more than 50 data points for each company, ranging from company location, to company industry sector such as health care, industrials, and energy, to market data such as market cap, and to financial data including revenue, income, inventory. Chapter 2.6.1 provides more detail on this process. The infrastructure we built is capable of pulling any of the other hundreds of data points Capital IQ gives access to historically, opening up the supply chain and strategic alliance networks to all Capital IQ's company data. This capability opens up a wide range of research opportunities for exploration.

2.4 Results

2.4.1 Static Structure. One area of particular interest regarding the global supply chain network is whether or not it lives up to its name of being a chain-like structure. We tested this concept by computing several static network statistics and analyzed their significance. We also computed several static statistics for the strategic alliance graph as well to better understand its structure.

2.4.1.1 Global Statistics. We began by analyzing the supply chain and strategic alliance networks on a global scale. This allowed us to look for large trends that may be affecting the supply chain overall as well as better understand the structure of the networks. These statistics can be found in Table 2.1.

Both networks have large diameters, adding weight to the idea of these networks having chainlike structure. Small world networks, such as the worldwide social network, in which six degrees of separation is believed to be the average distance between any two people, have the small average shortest path lengths between nodes [44]. The supply chain network has a 5.91 average shortest path length, while the strategic alliance network has a 6.48 average

Global Network Statistics		
Global Statistic	Supply Chain	Strategic Alliance
Diameter	18	22
Average shortest path	5.91	6.48
Clustering coefficient	.005	.017
Reciprocity	.09	n/a
Power law coefficient	2.76	3.12
Average degree	4.70	2.62
Median degree	2.0	1.0
Median indegree	1.0	n/a
Median outdegree	1.0	n/a

Table 2.1: A look into several global network values. The power law coefficient, which represents the power law degree fit of the degree distribution, is used to classify networks as scale-free. Generally, a power law coefficient between two and three qualifies a network as scale-free, hence the supply chain network meets this requirement while the strategic alliance network does not. However, this is not a firm requirement and 3.12 is close to three [8]. Thus each network may be viewed as scale-free. The supply chain network had a notably high reciprocity rate of 9%, a statistic we examine later in Figure 2.4. Values that require a connected graph were calculated on the largest weakly connected component for the supply chain network and on the largest connected component for the strategic alliance network. See Table 2.2 for largest component sizes.

shortest path length. A typical cutoff to qualify as a small world network is an average shortest path length proportional to the natural logarithm of the number of nodes in the network. The natural logarithm of the number of nodes in our networks is 5.25 for the supply chain network and 4.84 for the strategic alliance network. Hence neither network strictly meets this average shortest path length qualification, which is not a firm qualification and is open to interpretation, but the networks are not far off from it either. However, each network does have a small clustering coefficient, leading us to conclude these networks are not classifiable as small world networks.

Supply Chain Reciprocity. The reciprocity of the supply chain network is a key finding from our global statistical analysis. The reciprocity rate defines the rate at with a supplier customer relationship goes both ways, as in the customer also supplies the supplier. At .09, an entire 9% of company supplier relationships are mutual in this way. Compared to a configuration model graph based on the degree distribution of the supply chain graph, any rate over .03% would be considered an outlier. This reciprocity rate gives evidence that the

flow of goods in the worldwide supply chain is a complex process and that commonly used terms such us up stream and down stream supply chain may not fit as well as previously thought.

Further detail into this phenomenon is seen in the triad census of this network, which we see in Figure 2.1. Relationships involving reciprocity are much more likely in the supply chain network than in configuration model graphs produced by the degree distribution of the network. Specifically noting the elevated levels of triad 102, we see how far from expected reciprocation is in this network. Triads that may be expected to have an outsized role in the network's composition do not. One of these is 012, the triad containing one directed edge, and another is 012C, the triad containing two directed edges that flow from node 1 to node 3, which most closely models typical chain format. Thus the whole idea of viewing the supply chain as a chain for flow of goods is not so simple, and ignores important complexities and features found in the real network.

Degree Distributions. Another area of interest has been the degree distributions. These are given in Figure 2.3. The strategic alliance network has a large share of companies that only have one strategic alliance. This may explain in part the large diameter of the graph, and helps create a visual of its high alpha value as a distribution. Five times the number of companies have only one strategic alliance compared to those that have two alliances, the next most common amount. In comparison to the supply chain network, companies are only twice as likely to have only one connection when compared to two connections, which is also the second most common number of connections.

The indegree and outdegree histograms of the supply chain network give detail not gained from the overall degree distribution histogram. It is common for companies to not have any listed suppliers, or any listed customers. Capital IQ retrieves edge information from reports such as public companies financials and private company data. Companies that only have customers or suppliers are likely private companies, which are not required to release customer and supplier reports as public companies are. Thus these companies with

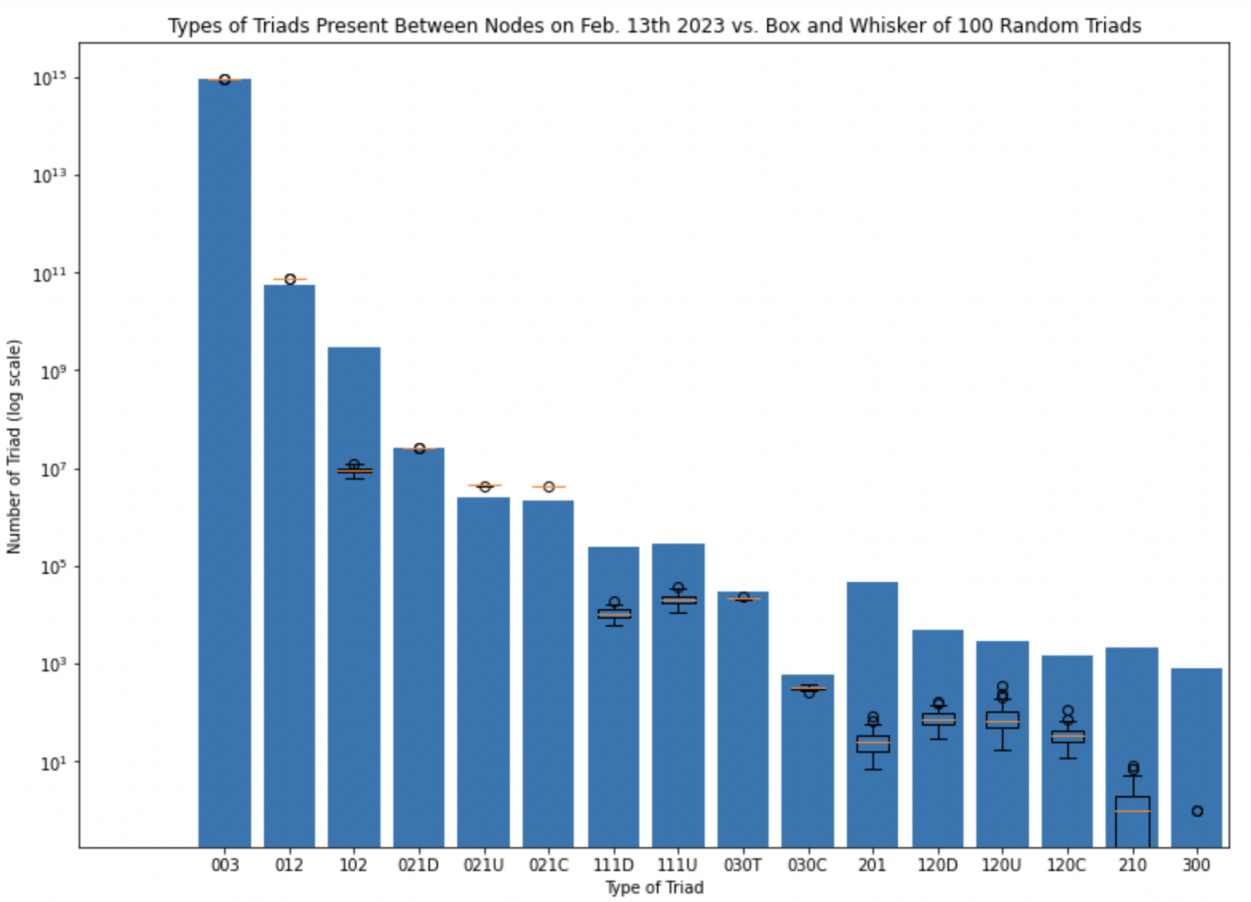

Figure 2.1: Triadic census of the supply chain network. Values for one hundred configuration model graphs produced by the degree distribution of the supply chain network degree distribution are given in box and whisker plots, and the values for the supply chain network are given by the bar chart. Triad 102 denotes reciprocation, which is much more common than what is seen in the configuration models. Triads 111D, 111U, 201, 120D, 120U, and 120C involve one edge reciprocated with the other edges not, and triads 201, 210, and 300 involve two of the three edges reciprocated. See Figure 2.2 for definitions of each triad type.

only suppliers or customers may have been listed on other companies' required or voluntary reports. All of these companies with either only customers or suppliers, roughly $130,000$ in all, or 73% of the network, can only be the beginning or end of supply chain subnetworks found in the data. Another point of interest here is that companies are more likely to have one supplier than they are to have none, and there are more companies in indegree than outdegree for degree counts three through 21. Thus the outdegree is dominated by few companies that supply many other companies, and the indegree has a more evenly distributed spread of how many suppliers each company has.

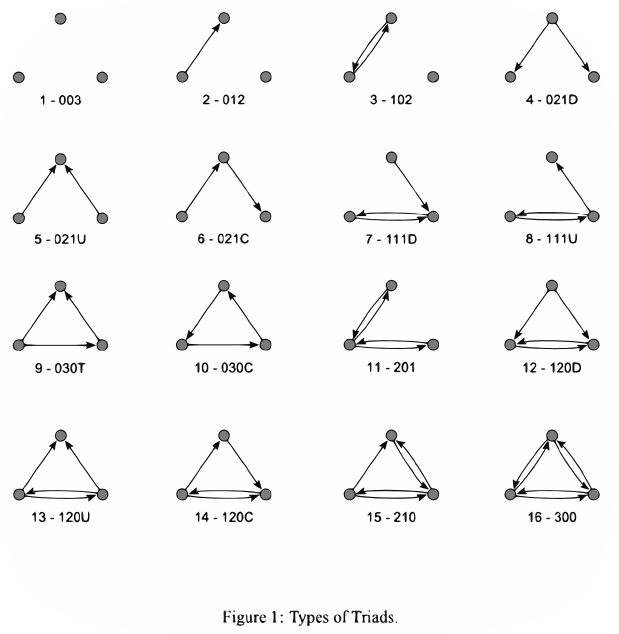

Figure 2.2: Triad definitions, taken from [3].

Connected Component Statistics and Applications. Much of the statistics and tests we run on these networks require using a connected component. For these we use either the largest weakly connected component for the supply chain network, or the largest connected component for the strategic alliance network. The sizes of these components are given in Table 2.2. These giant components consist of the vast majority of companies in each network, hence statistics computed on these components are nearly global. This also demonstrates that at the global level, there are not multiple distinct supply chains competing with each other for dominance, there is only one.

The supply chain network's largest strongly connected component takes up over 10% of the network, creating a strong center to our network with increased flow of goods and services. The largest weakly connected component increases this view of connectedness of the network, with over 90% of companies connected when viewed as an undirected graph.

2.4.1.2 Centrality Measures. A major strength of network analysis is centrality measures, which help us define various measures of importance. Traditional views of company importance may rely solely on financial data such as revenue or market data like stock prices.

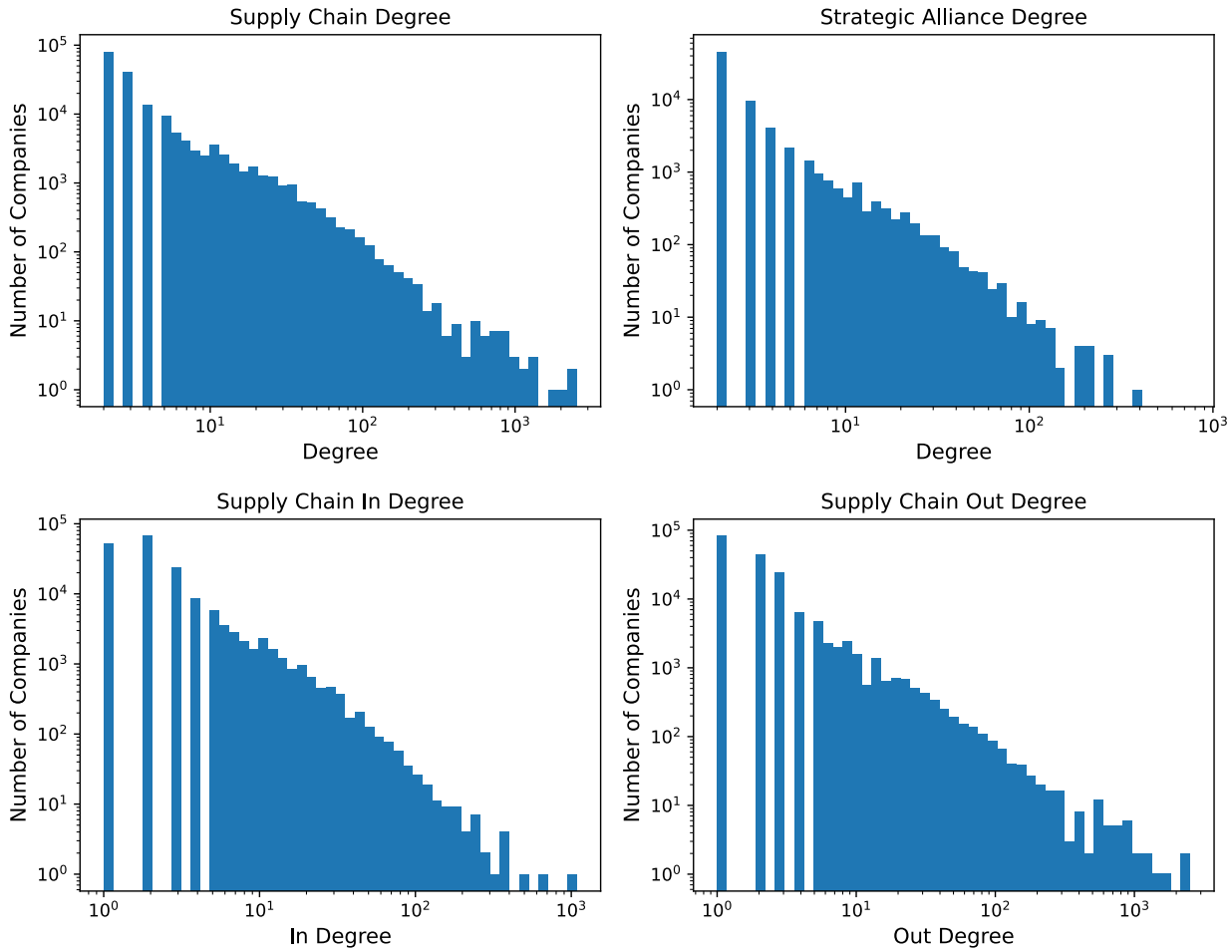

Figure 2.3: Histograms containing the degree distributions of both networks plotted on loglog scale to demonstrate how each networks' distributions follow a power law. This gives us a visual of the higher power law coefficient for the strategic alliance network given in Table 2.1 compared to the supply chain power law coefficient. The maximum degree in the supply chain network is more than $1,500$ and in the strategic alliance network the maximum degree is more than 700. Also note how it is more likely for a company to not have any listed suppliers than it is for the company to have no listed customers.

Centrality measures calculate importance various ways, measuring features from shortest average path to other companies in the graph for closeness centrality to being labeled as important by being connected to important companies with eigenvector centrality. These measurements provide greater insight into which companies are positioned most strongly within the supply chain and strategic alliance networks, which can be a good indicator of influence. The top ten companies for various centrality measures for each network are given in Table 2.3 and Table 2.4.

Connected Component Sizes for Both Networks		
Component	Supply Chain	Strategic Alliance
Largest connected component (Percentage of network)	n/a	50745 (73.2%)
Second largest connected component (Percentage of network)	n/a	39 (.06%)
Largest weakly connected component (Percentage of network)	164939 (92.7%)	n/a
Second largest weakly connected component (Percentage of network)	82 (.05%)	n/a
Largest strongly connected component (Percentage of network)	20327 (11.4%)	n/a
Second largest strongly connected component (Percentage of network)	33 (.02%)	n/a

Table 2.2: Largest and second-largest connected component sizes for each network. The vast majority of each network can be found in their giant component using the weakly connected component for the supply chain network. In particular, the supply chain network is extremely well-connected with over 90% of companies residing in the largest weakly connected component.

Across both networks, technology companies such as Microsoft had high centrality values. This gives evidence of these companies having strong supply chain networks and being connected to other important companies on various projects through strategic alliances. A difference between these networks is that banks hold a strong position in the supply chain network, while they do not in the strategic alliance network. Difference might be attributed to their tendency not to branch outside of the financial sphere, leading them to keep all their work under their own company while avoiding strategic alliances. Most companies work with a financial institution or bank for auditing and consulting, which is the main driver behind banks being well-connected across the supply chain network.

2.4.1.3 Edge Label Statistics, Definitions, and Significance. Edge labels giving the type of business relationship between two companies are a unique qualifier in our dataset for analysis. In total our dataset contains eight edge labels, as shown in Table 2.5. These edge labels each presents measures of collaboration between two businesses, and can measure the business impacts of broken relationships. These edges also provide a look at supply chain subsets, such as the worldwide creditor network and worldwide landlord network. Giving

Centrality Measures for Supply Chain Network	
Centrality Measure	Supply Chain Top 10
Degree centrality (company degree)	1. Bank of America (2545) 2. Wells Fargo Bank (2412) 3. JPMorgan Chase Bank (2191) 4. PNC Bank (1727) 5. U.S. Bank (1388) 6. Citibank (1322) 7. Amazon.com (1271) 8. Truist Bank (1144) 9. Link Intime India Pvt. Ltd. (1096) 10. Continental Stock Transfer and Trust Company (1026)
Indegree centrality (company indegree)	1. Amazon.com (1087) 2. Walmart Inc. (469) 3. Apple Inc., (469) 4. Renfe Operadora S.C. (387) 5. Adore Beauty Pty Ltd (384) 6. Auto Partner SA (376) 7. PT Metrodata Electronics Tbk (354) 8. Dufry AG (341) 9. Samsung Electronics Co., Ltd. (262) 10. Alphabet Inc. (262)
Outdegree centrality (company outdegree)	1. Bank of America (2527) 2. Wells Fargo Bank (2395) 3. JPMorgan Chase Bank (2172) 4. PNC Bank (1719) 5. U.S. Bank (1383) 6. Citibank (1303) 7. Truist Bank (1139) 8. Link Intime India Pvt. Ltd. (1095) 9. Continental Stock Transfer and Trust Company (1026) 10. Goldman Sachs Bank USA (953)
Betweenness centrality	1. Amazon.com 2. Microsoft Corporation 3. Apple Inc. 4. ICICI Bank Limited 5. Repsol, S.A. 6. Walmart Inc. 7. Amazon Web Services, Inc. 8. Google LLC 9. CT Real Estate Investment Trust 10. Alphabet Inc.
Closeness centrality	1. Amazon.com 2. Citibank 3. JPMorgan Chase Bank 4. Bank of America 5. BNP Paribas SA 6. Wells Fargo Bank 7. MUFG Bank 8. IBM Corporation 9. Mastercard Incorporated 10. Microsoft Corporation
Eigenvector centrality	1. Amazon.com 2. Apple Inc. 3. Walmart Inc. 4. Microsoft Corporation 5. TD Synnex Corporation 6. Synnex (Thailand) Public Company Limited 7. Dicker Data Limited 8. Softlogic Holdings PLC 9. CDW Corporation 10. Dustin Group AB

Table 2.3: Top 10 companies for various centrality measures for the supply chain network. Generally technology companies and banks tended to have higher centralities. In particular, banks have high outdegree centrality due to having a large number of customers, while technology companies have higher indegree centralities due to having more suppliers. Betweenness centrality, a measure of being on the shortest path through the network, closeness centrality, a measure of distance to other companies in the network, and eigenvector centrality, a measure of being connected to important companies, tended to favor technology companies. Technology companies may tend to have stronger supply chains than companies do in other industry sectors.

Centrality Measures for Strategic Alliance Network	
Centrality Measure	Strategic Alliance Top 10
Degree centrality (company degree)	1. Microsoft Corporation (753) 2. IBM Corporation (368) 3. Oracle Corporation (278) 4. Cisco Systems, Inc. (274) 5. Amazon Web Services, Inc. (259) 6. Alphabet Inc. (221) 7. Amazon.com (215) 8. Samsung Electronics Co., Ltd. 9. Google LLC (205) 10. HP Inc.
Betweenness centrality	1. Mahindra Lifespace Developers Limited 2. Cisco Systems, Inc. 3. Pintrest, Inc. 4. Crinetics Pharmaceuticals, Inc. 5. Hansfort Investment Pte Ltd 6. GABO STAHL GmbH 7. SAP Asia Pte. Ltd. 8. Anywhere Real Estate Inc. 9. IKEA B.V. 10. TAG Colonia-Immobilien AG
Closeness centrality	1. Mahindra Lifespace Developers Limited 2. Cisco Systems, Inc. 3. TAG Colonia-Immobilien AG 4. Kauflandervice GmbH and Co. KG 5. Crinetics Pharmaceuticals, Inc. 6. GABO STAHL GmbH 7. Hansfort Investment Pte Ltd 8. Pinterest, Inc. 9. Shop Rite, Inc. 10. Singapore Telecommunications Limited
Eigenvector centrality	1. Microsoft Corporation 2. IBM Corporation 3. Oracle Corporation 4. Cisco Systems, Inc. 5. VMware, Inc. 6. Amazon Web Services, Inc. 7. SAP SE 8. HP Inc. 9. Intel Corporation 10. Alphabet Inc.

Table 2.4: Top 10 companies for various centrality measures for the strategic alliance network. U.S. Technology companies generally have the highest number of strategic alliances, seen under degree centrality as well as the highest eigenvector centrality, a measure of being connected to important companies. Betweenness centrality, a measure of being on the shortest path between other companies in the network, and closeness centrality, a measure of having the shortest paths to other companies, were more global in scope. These measures have several companies from Asia and Europe while retaining a small proportion from the United States.

unparalleled detail into the strengths of these networks, each of these eight networks could be also analyzed using this chapter's methods, such as centrality measures and community detection methods.

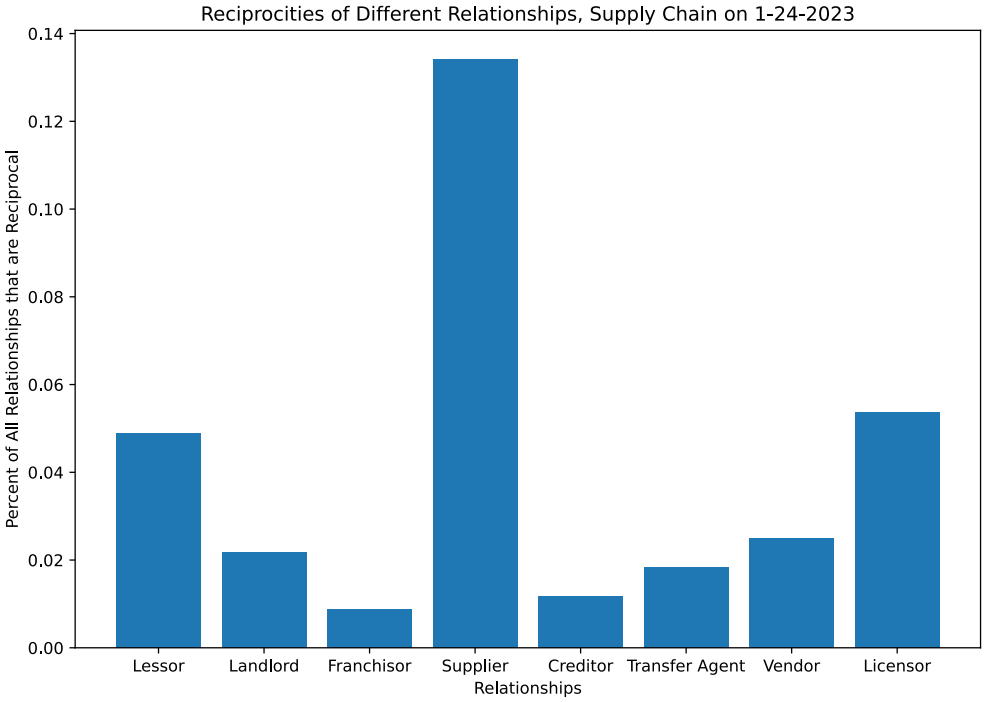

Figure 2.4: The edge labels allow us to better understand global trends within the supply chain network such as the high reciprocity rate given in Table 2.1. Here we see this rate is nearly all driven by supplier relationships, indicating companies often pass products back and forth to each other for different stages of the production process.

Demonstrating these further insights, Figure 2.4 presents a deeper investigation into the high reciprocity rate seen in Table 2.1 through the edge labels lens.

In this case, the edge labels show when a business relationship has a high chance of reciprocity. The data shows that supplier relationships are the predominant type with high odds of reciprocity. This data also matches well with intuition that a majority of low-reciprocity relationships need uncommon circumstances to occur, such as a creditor being the creditor to their creditor.

2.4.1.4 Static Value Chain Business Results. The static view of the global value chain provides a variety of business applications. One of these business applications is

Types of Supplier-Customer Relationships		
Relationship	Definition	Number of Relationships (Percentage of Network)
Vendor	Company providing incidental services that does not form part of the main services to the subject company. For example: A company providing transport services to an IT services company is a vendor.	51291 (12.2%)
Transfer Agent	Records changes of ownership, maintain the issuer's security holder records, cancel and issue certificates, and distribute dividends.	76882 (18.3%)
Supplier	An entity which supplies goods, services and products to the subject company for a consideration.	141002 (33.6%)
Creditor	An entity (public or private group, or a financial institution) that makes funds available to another with the expectation that the funds will be repaid. Repayment will include the payment of any interest or fees.	110037 (26.2%)
Lessor	An entity that provides the right to use an asset for an agreed period of time in return for a payment or series of payments to the lessee.	8945 (2.1%)
Landlord	An entity that provides the right to use an asset for an agreed period of time in return for a payment or series of payments to the lessee.	20089 (4.8%)
Franchisor	A franchise is an investment in which you pay another business for the right to use its business model and products. A franchisor is the party granting the franchise right to the subject company.	1079 (.3%)
Licensor	An entity which grants a license to a company.	10212 (2.4%)

Table 2.5: Definition of each relationship type in the supply chain network and its percentage of total relationships in the network as of 2/20/23. These relationship types add value to our graph by defining what kind of business interaction two companies have and allowing us to view subnetworks of the data by restricting to certain relationship labels. Definitions are from Capital IQ [14]. Four of eight relationship types, vendor, transfer agent, supplier, or creditor relationships make up over 90% of the data.

understanding the importance of the centrality measures. Further research into the direct impact of each centrality measure would be necessary to make conclusive connections between each measure and business results, but in Section 2.2.3 we outlined previous research that found that diversifying a company's suppliers within the supply chain affects a company's overall robustness. Thus businesses that use centrality measures when choosing new suppliers may be able to directly impact the robustness of their supply chain. They could accomplish this by choosing important companies to work with, which would diversify their supply chain within the overall supply chain network.

Another business application from our results stems from the near small world average shortest path length size inside the supply chain network. This small average shortest path length adds understanding to why shock propagation can spread quickly through a supply chain network as discussed in Section 2.2.3. Businesses that are aware that they may be closely related to any company within the global supply chain network can analyze what areas of their supply chain will be adversely affected from shock propagation when another company in the global supply chain is dissolved. These businesses can then act on this data to increase supply chain redundancies before the effects from shock propagation reach them, bettering their business situation.

2.4.2 Dynamic Structure. The dynamic view of the supply chain and strategic alliance networks is a key for identifying global value chain trends and risks. We began consistent clean data pulls on January 20th, 2023, and have since continued near daily pulls. Due to Capital IQ's policies, a relationship is listed if it has been reported anytime in the last three years. Thus any relationships that has dropped out of either network has been inactive for up to three years. To begin our analysis, we tracked the daily change rate. Table 2.6 shows the first month's average daily change.

While the daily change rate is minimal, daily data collection is crucial to a dynamic network analysis; and these small fluctuations could be useful in early discovery of future long-term trends.

Network Average Daily Change Over One Month		
Network	Supply Chain	Strategic Alliance
Average number of companies in the previous day's data but not the current day (percentage of overall network)	74 (.04%)	5 (.001%)
Average number of companies in the current day's data but not the previous day (percentage of overall network)	116 (.07%)	40 (.06%)
Average number of companies in either the current or the previous day's data but not both (percentage of overall network)	190 (.10%)	45 (.07%)
Average number of relationships in the previous day's data but not the current day (percentage of overall network)	285 (.07%)	12 (.01%)
Average number of relationships in the current day's data but not the previous day (percentage of overall network)	497 (.12%)	64 (.07%)
Average number of relationships in either the current or the previous day's data but not both (percentage of overall network)	782 (.19%)	76 (.09%)

Table 2.6: Average daily change in supply chain and strategic alliance networks over the first month of data collection. Each network has experienced consistent daily growth, gaining new companies and density. The supply chain dataset has been particularly dynamic, changing at twice the rate of the strategic alliance network. The strategic alliance network has significantly lower company and relationship dropout rates than the supply chain dataset, indicating a strategic alliance relationship could have higher longevity than relationships in the supply chain network. Company dropout is reported with a three-year lag.

In Table 2.7 we tracked how the networks changed over the first month of data collection. We observed strong trends of network growth and minimal company dropout. The supply chain network is much more volatile than the strategic alliance network, with more than five percent of relationships either formed or dissolved during the first month. Similarly, over three percent of companies either joined or left the supply chain over the first month.

The final aspect of the dynamic supply chain is the company statistics included in the dataset. We will pull this data quarterly to align with company quarterly financial reports. The data includes more than 50 features, ranging from categorical data such as industry and country to quantitative data such as inventory, revenue, and tax rates. Since the dataset is new, there has only been one financial pull up to this point.

Network Change Over One Month		
Network	Supply Chain	Strategic Alliance
Total companies in network as of 1/20/23	176721	68280
Total companies in network as of 2/20/23	177999	69325
Companies in network on 1/20/23 but not on 2/20/23 (percentage of overall network)	2207 (1.3%)	146 (.2%)
Companies in network on 2/20/23 but not on 1/20/23 (percentage of overall network)	3485 (2.0%)	1191 (1.7%)
Companies in either 1/20/23 or 2/20/23 but not both (percentage of overall network)	5692 (3.2%)	1337 (1.9%)
Total relationships in network as of 1/20/23	413207	89272
Total relationships in network as of 2/20/23	419537	90854
Relationships in network on 1/20/23 but not on 2/20/23 (percentage of overall network)	8568 (2.1%)	346 (.4%)
Relationships in network on 2/20/23 but not on 1/20/23 (percentage of overall network)	14898 (3.6%)	1928 (2.1%)
Relationships in either 1/20/23 or 2/20/23 but not both (percentage of overall network)	23466 (5.6%)	2274 (2.5%)

Table 2.7: Change in supply chain and strategic alliance networks after the first month of data collection. The change has been significant, especially in the supply chain network. While the strategic alliance network hardly deals with companies dropping out of the network, the supply chain network receives and loses a significant number of companies. This indicates the importance of analyzing the worldwide supply chain in the dynamic sense, as a static view can quickly become dated. Company dropout is reported with a three-year lag.

2.4.2.1 Dynamic Value Chain Business Results. A significant finding from our dynamic global value chain analysis is how quickly the value chain is changing, particularly the global supply chain section of the value chain. The supply chain is incredibly dynamic, with only a month's observation seeing over 5% of relationships either being created or being dropped, with a three-year lag for dropped relationships. This high rate of change is similar to what we would expect from resilient supply chains, which are supply chains that adapt and create new relationships to avoid disruptions as discussed in Section 2.2.3. This knowledge of the shifting business environment is also useful for business leaders; realignment through new companies and relationship formation for other companies presents new opportunities for companies, such as different suppliers or backup suppliers to ensure supply chain redundancy. Switching to different suppliers might increase quality or reduce costs for a business, and from Section 2.2.3 we find that supply chain redundancy is crucial to resist shock propagation

from disasters or demand shocks.

2.4.3 Community Structure. For future research, network community structure provides an area of interest. Investigating community structures could provide insight into how rival companies create their supply chains, which industry sectors have the highest collaboration rates, and larger trends such as supply chain reshoring [38]. The communities we explore in this section are not unique; they are dependent on the community detection algorithms and parameters we chose to implement and re-running the algorithms returns slightly different communities. Major community structures and findings presented in this section were similar when re-running the algorithms, while the exact percentages of community attribute makeups presented in various tables throughout this section varied slightly.

2.4.3.1 Global Community Detection. Our first step in community detection and analysis was verifying that community structure significantly differed from a similar but random graph. We created one hundred configuration model random graphs based on the supply chain network degree distribution. We then compared these graphs' community sizes and partition modularity to those of the original supply chain and strategic alliance graphs. We used two community detection methods, the Reichardt and Bornholdt's Potts Model (RB) and the Constant Potts Model (CPM). Figure 2.5 compares these two community detection methods applied to the supply chain against the configuration model. Each method finds significantly higher partition modularity for the supply chain network, signaling community structure beyond what can be found inside a random graph of this size and structure. The RB method resulted in a partition with over .5 higher modularity for the majority of resolution parameters for partitions of the supply chain network in comparison to partitions of the configuration model network. The RB method is trained on optimizing modularity and the CPM method is optimized on maximizing internal edges in communities while maintaining a small community size. Thus it is not surprising the RB method outperformed CPM according to modularity scores. Figure 2.6 shows the same comparison applied to the strategic alliance network. The RB method again found higher modularity partitions for the strategic

alliance network compared to the configuration model's partitions. The CPM method found marginally higher modularity for the strategic alliance graph when compared to the configuration models, and only when using low resolution parameter values. Thus each community model on the supply chain and strategic alliance networks gives evidence that the networks contain significant community structure. Due to the RB method maximizing modularity, creating a measurable distinction between the communities found in each network and the communities of their configuration models, we will use this method to analyze community structure throughout this section. These methods of community detection are discussed in more detail in Section 2.6.2.

Global Community Breakdown by Industry Sector. We now take a further look into the communities discovered by the RB method. We analyzed the largest 12 supply chain network communities created by this method at a resolution parameter of .7. We used company industry sector labels to break down the partitions. This is just one of the more than 50 statistics we have for each company in this network. There are 12 distinct industry sectors and in all, over 70% of companies have their industry sector listed. From these top communities, we sorted which ones are significant in terms of their company's industry sector. We measured significance by comparing the industry sectors' representation percentages in the overall network to their representation percentages within each community and by measuring what percentage of the companies in each industry sector ended up in a community together. Table 2.8 shows these results.

Eight of the twelve largest communities proved significant by these measures. Interesting groupings included a community formed mostly through information technology and communication companies, another through consumer discretionary and consumer staples companies, and another with a majority makeup in health care companies. This community detection gives us more understanding of which industry sectors are most likely to work together, and of which industries are least likely to work together. By tracking communities over time, we can watch for community realignment, such as when certain industry sectors

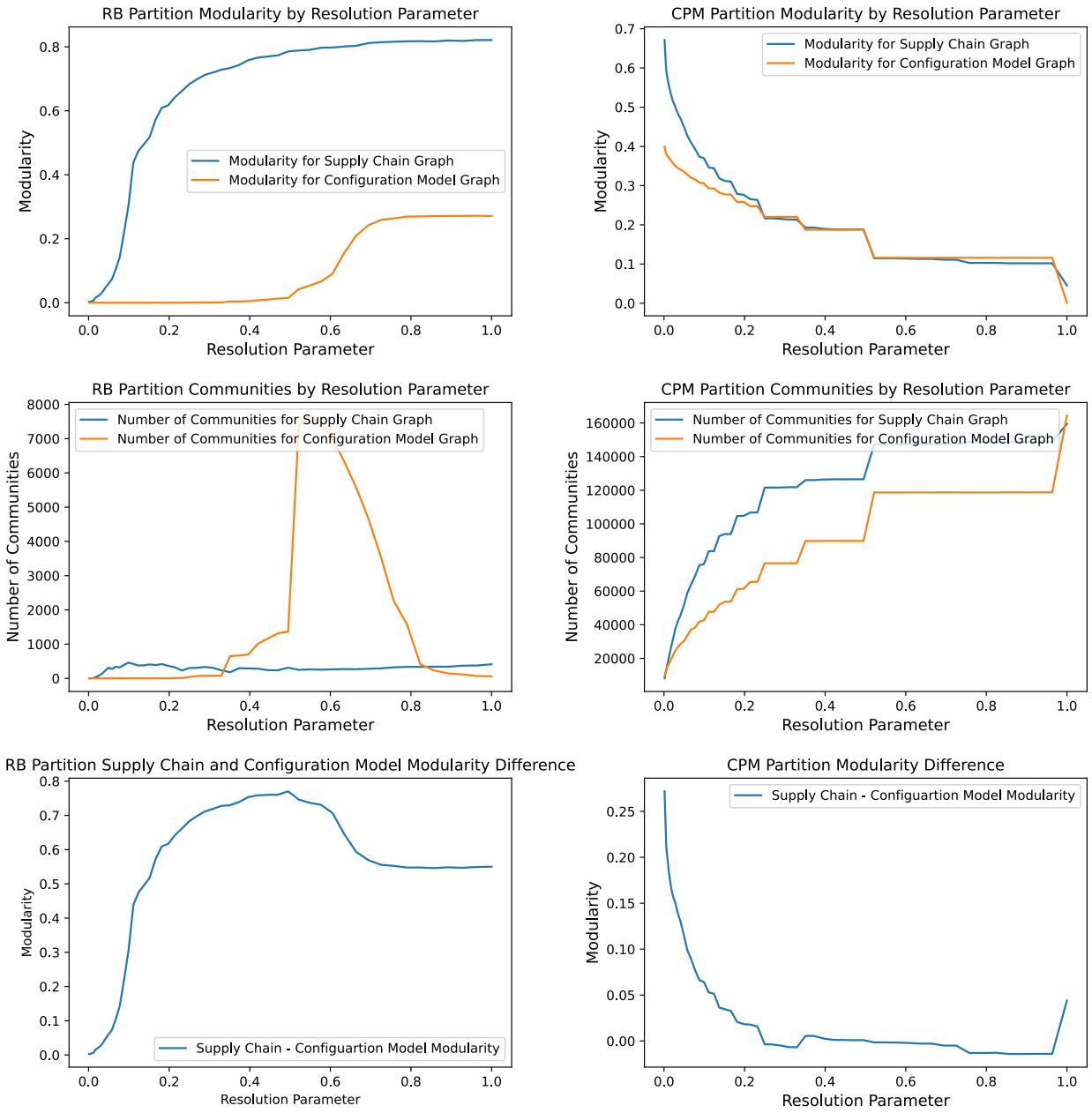

Figure 2.5: Partition quality comparison between the supply chain network and configuration model based on the degree distribution of the supply chain network. The community detection methods used here are the Reichardt and Bornholdt's Potts Model (RB) and the Constant Potts Model (CPM). Here both methods display significant community difference between the configuration model and the supply chain. The RB method finds partitions of much higher modularity and keeps the number of communities found fairly consistent across all resolution parameters. The CPM method creates higher modularity partitions for the supply chain with low resolution parameters but quickly descends to match the configuration model graph. For all resolution parameters this methods finds many more communities than the configuration model. Both these methods together indicate significant community structure exists within the supply chain network.

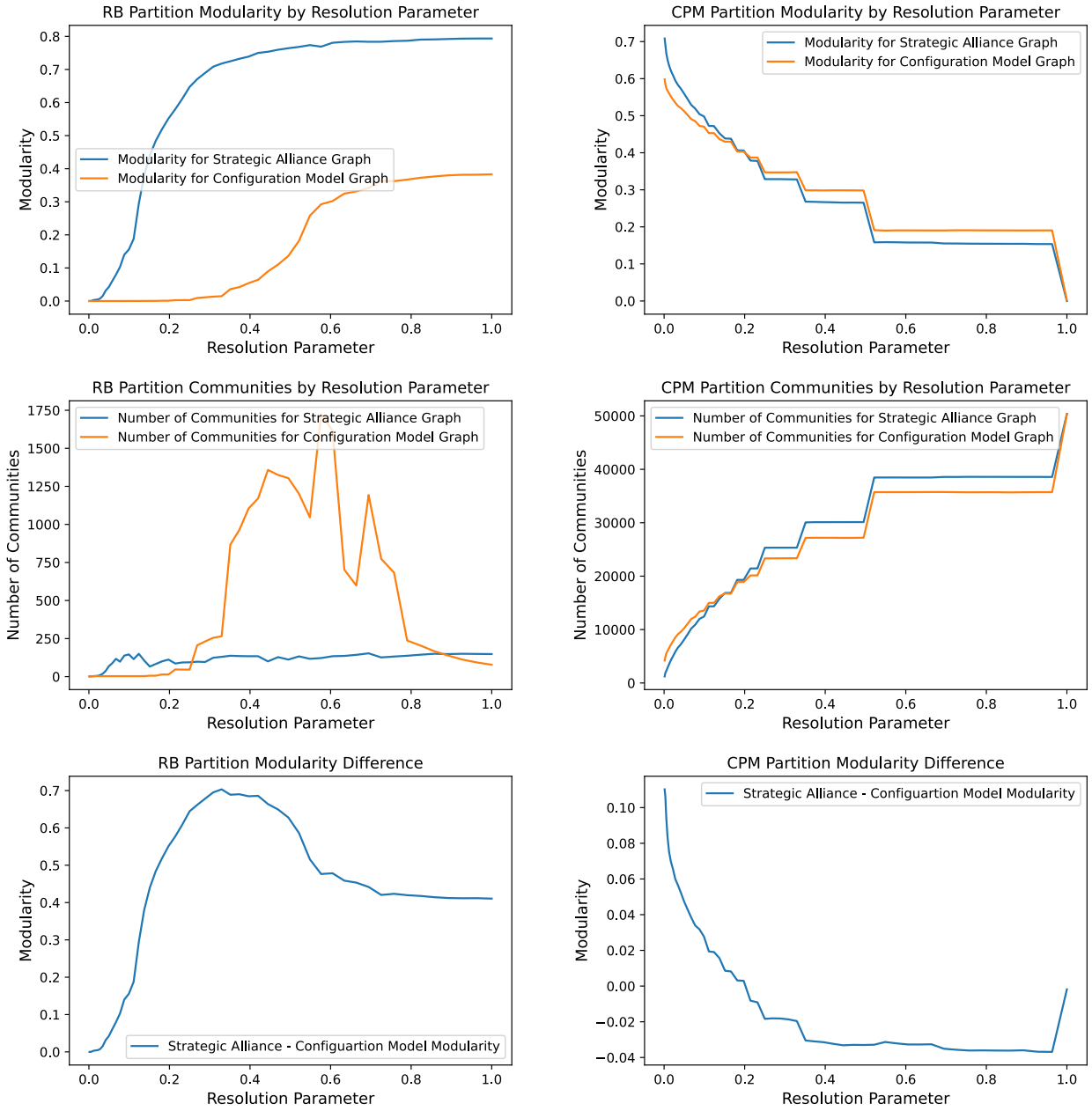

Figure 2.6: Partition quality comparison between the strategic alliance network and configuration model based on the degree distribution of the strategic alliance network. The community detection methods used here are the Reichardt and Bornholdt's Potts Model (RB) and the Constant Potts Model (CPM). Here the CPM method only briefly differed from what we found with the configuration model. The RB method discovered high modularity partitions in this network, while maintaining a consistent number of communities across resolution parameters.

Significant Global Communities by Industry Sector	
Community ID	Top Industry Sectors as Percentage of Community (Percentage of Community, Percentage of Overrepresentation Compared to Overall Network)
2	1. Financials (19.5%, 71.0%) 2. Energy (8.0%, 207.3%) 3. Utilities (4.6%, 85.9%)
3	1. Information Technology (27.6%, 297.0%) 2. Communication Services (19.8%, 475.8%)
6	1. Consumer Discretionary (28.1%, 156.9%) 2. Industrials (25.7%, 69.8%)
7	1. Health Care (54.5%, 832.3%)
8	1. Consumer Discretionary (39.8%, 264.1%) 2. Consumer Staples (14.9%, 256.9%)
9	1. Information Technology (22.6%, 232.9%)
10	1. Materials (19.1%, 215.0%)
11	1. Industrials (25.1%, 65.9%) 2. Energy (12.5%, 379.2%)

Table 2.8: Selection of significant communities by the 12 industry sector classifications. Community number relates to the size of the community, with 1 being the largest community found in the network. Here communities were detected through the RB method with a resolution parameter of .7. This returned more than 400 communities, of which we analyzed the largest 12. Health care proved to have the strongest community structure of any of the classifications, with almost half of the companies in the same community. Our cutoff for listing an industry sector in the table was at least 10% of the community being from that industry sector. If an industry sector met either of these requirements it was listed in both columns.

grow closer or further apart.

2.4.3.2 Health Care Industry Sector Community Analysis by Industry. While understanding communities at a global level is key for understanding the structure of the worldwide supply chain, there is also a lot to be gained through analyzing these communities individually, or by finding communities by industry sector. Due to the strong global community found in health care, we break down the companies in this industry sector to see if we can learn more about how the companies within this industry sector operate and compete.

For each company, we have four levels of industry information. From broadest to most specific these are: industry sector, industry group, industry, and primary industry. The health care industry sector has $9,733$ companies, divided into two different industry groups, six different industries, and 10 different primary industries. Due to the small sizes of some of these groups, we used the six different industry classifications to gain more insight into health care companies. These are life sciences tools and services, biotechnology, health care providers and services, pharmaceuticals, health care equipment and supplies, and health care technology.

Health Care Community Detection Methods. First, we looked into Global Community 7 found in Table 2.8. For each industry we computed its share of health care within Community 7 as well as its representation compared to the industry's overall representation within health care. Table 2.9 shows this data.

Our analysis shows that Community 7, the large global health care community, is built on strong representation of biotechnology and pharmaceutical industries. Thus these industries likely have some of the closest ties for any two industries within the global supply chain.

After this analysis, we delved deeper into the global health care industry sector of the supply chain network. Consisting of $9,733$ companies, this subnetwork had one weakly connected component of $5,293$ companies, with the remaining companies disconnected. This disconnectedness was due to us only using companies in the health care industry sector, many

Global Community 7 Health Care Companies by Industry		
Industry	Industry percentage of Health Care within Community 7	Industry Over/Under Representation Compared to Representation within Health Care
Biotechnology	33.2%	68.2%
Pharmaceuticals	29.1%	12.1%
Health Care Providers and Services	16.4%	-44.9%
Health Care Equipment and Supplies	10.4%	-26.1%
Life Sciences Tools and Services	8.7%	42.5%
Health Care Technology	2.2%	-49.6%

Table 2.9: Health care industry sector break down by industry within global Community 7. Biotechnology makes up the largest industry in this community as a percentage of health care companies. Biotechnology also has the highest overrepresentation rate of the six health care industries. Health care providers and services as well as health care technology are the most underrepresented industries in health care within Community 7.

of which only had supplier and customer connections outside of the industry sector. While this weakly connected component contained 54.4% of all health care firms, it contained 84.2% of the health care firms contained in Global Community 7. Thus we used this component as our subnetwork to analyze the global health care community. We divided this subnetwork into 50 to 65 communities using various resolution parameters and the RB method. These partitions generally had a modularity score of .77, where partitions generated on a configuration model of the graph had a modularity score of .53, indicating the community structure in the health care industry sector is better than expected from a random graph.

Health Care Community Detection Results. This led to interesting results regarding competition within the health care industry. Table 2.10 summarizes some of the largest communities and their respective major industries. A strong trend throughout the communities was that a large number of them contained a large proportion of pharmaceutical and biotechnology firms working together. This is consistent from what we found in the global health care community. A difference however, is that while the global community analysis gave insight on how separate health care companies tend to be from the rest of

Health Care Communities by Top Industries	
Community ID	Top Industries as Percentage of Community (Percentage of Community, Percentage of Over/Under Representation Compared to Representation within Health Care)
1	1. Pharmaceuticals (35.6%, 12.7%) 2. Biotechnology (29.9%, 10.6%) 3. Health Care Providers and Services (19.6%, -2.9%)
2	1. Biotechnology (44.3%, 64.1%) 2. Pharmaceuticals (38.5%, 21.6 %) 3. Life Sciences Tools and Services (8.1%, .5%)
3	1. Biotechnology (58.3%, 115.8%) 2. Pharmaceuticals (20.3%, -35.8%) 3. Life Sciences Tools and Services (12.5%, 54.2%)
4	1. Pharmaceuticals (70.5%, 122.8%) 2. Health Care Providers and Services (19.9%, -1.6%)
5	1. Biotechnology (52.0%, 92.3%) 2. Health Care Providers and Services (20.0%, -1.1%) 3. Life Sciences Tools and Services (8.6%, 5.6%)
6	1. Pharmaceuticals (73.4%, 132.1%) 2. Health Care Providers and Services (11.0%, -45.7%) 3. Biotechnology (10.4%, -61.5%)
7	1. Biotechnology (28.7%, 6.0%) 2. Life Sciences Tools and Services (25.0%, 208.4%) 3. Health Care Equipment and Supplies (18.9%, 82.9%) 4. Health Care Providers and Services (17.1%, -15.5%)
8	1. Biotechnology (37.6%, 39.3%) 2. Pharmaceuticals (37.7%, 19.1%) 3. Health Care Providers and Services (10.5%, -15.5%)
9	1. Pharmaceuticals (54.1%, 70.9%) 2. Biotechnology (20.3%, -25.0%) 3. Health Care Providers and Services (16.2%, -19.8%)
10	1. Biotechnology (42.6%, 57.4%) 2. Pharmaceuticals (21.3%, -32.7%) 3. Health Care Providers and Services (14.9%, -26.3%) 4. Health Care Technology (7.8%, 190.8%)

Table 2.10: Top 10 health care industry sector communities by industry, ranked according to size. These were found through the RB method with a resolution parameter of 1. Notable is a pattern of biotechnology and pharmaceutical firms forming strong communities together, mirroring what we saw in the global health care community. However, each of those industries also formed one community largely consisting of themselves with some health care providers and services companies. We used 7.5% community participation rate as the cutoff to list.

the supply chain, with the vast majority of the largest weakly connected component of the health care industry sector residing in this community, our analysis of the community structure in the health care sector gave insight on the competition within this sector. Far from being united, we learned that pharmaceutical and biotechnology firms have created several competing community structures amongst themselves, with six of the 10 largest communities being dominated by these two industries. While these industries were the second and third largest industries within health care, in each of these six communities where they constituted the top two industries by representation, at least one of these industries was vastly overrepresented.

Another interesting find in the community structure was that the fourth largest community was dominated by pharmaceutical firms with very low biotechnology firm representation and that the fifth largest community was dominated by biotechnology firms with very low pharmaceutical firm representation. Thus, while these two industries tend to work together, each industry has a distinct segment that works with the other industries within health care while nearly avoiding the other industry. While these two industries were dominant in size, the other industries having consistent presence throughout the communities also has implications. These industries do not tend to join their supply chains together, indicating that they could consist of more support level firms needed in every supply chain community for functionality.

2.4.3.3 Community Structure Business Results. Even though the communities we found are not unique, they do reflect groupings of companies that have a high number of relationships within each community when compared to the number leaving the community. Companies interested in strengthening their supply chain robustness could look for suppliers outside of communities they reside in. In our research of the health care industry sector, we found that the majority of communities have large numbers of pharmaceutical or biotechnology firms as seen in Table 2.10. Businesses looking for new partners in either of these industries could increase their supply chain robustness and resilience by partnering

with companies belonging to different communities. This agrees with research we covered in Section 2.2.3, which found that companies that diversified their supply chains were more robust against shock propagation and other supply chain interrupting events. Though in this section we focused on global communities and health care communities, the same code we wrote for these findings can be applied to any industry.

2.5 Discussion

Foremost among our results is the utility of this dataset. We have shown how vital it is to view supply chains dynamically. Over the course of the month that we tracked the supply chain, we saw that over 5% of relationships in the chain had changed and over 3% of companies had changed. Thus static views of the global supply chain can quickly becomes an inaccurate description of the current environment. Furthermore, we demonstrated this dataset's depth and breadth, as it contains both the global supply chain network and the global strategic alliance network. In our literature review we were unable to find anything resembling a global strategic alliance network or a global supply chain network of a similar size and accessibility to ours.

Our company-to-company relationship data added significant depth to our supply chain dataset. We have eight supply chain network relationship labels that allow us to find deep insights into business relationships and global trends and statistics, as we did with the reciprocity in Figure 2.4. These labels aid study into the variability of business relationships, of which our study found may. Initially, we gathered supply chain network reciprocity levels, but found through our edge labels that suppliers are much more likely to have reciprocal relationships than any other relationship type.

We also demonstrated how company labels and statistics can be applied to analyze the global value chain network, which we did to measure company importance through centrality measures in Table 2.3 and Table 2.4 and through community detection in Section 2.4.3. Our centrality measure analysis was significant in identifying central companies to both the

supply chain network and the strategic alliance network. We found that in both networks technology companies have a strong position regarding several centrality measures and that banks are key to the supply chain network, but not the strategic alliance network. A major difference between the companies in the centrality rankings of the two networks was that the supply chain network had a higher proportion of companies from the United States. This may be due to the United States central role to the world's economy or companies in other areas of the world being more likely to form partnerships through strategic alliances.

As part of our community detection research we found significant communities globally. We found that health care is the most clustered industry sector globally. This has implications that the health care supply chain may be relatively isolated and operating more independently from other industry sectors. Our deeper analysis into the health care sector revealed strong competition among many clusters formed largely of biotechnology and pharmaceutical companies. Thus our research demonstrated how our dataset is useful in not only learning about general global trends, but also learning trends about specific communities.

While our dataset has significant improvements and advantages compared to other datasets, it does have several shortcomings. First, CapitalIQ has over $12,300,000$ firms in their database, and only supply chain relationship data for $175,000$ firms. Also, CapitalIQ reports relationships if they have been announced in a statement by either company any time in the last three years. This makes our dataset three years behind on dissolved relationships, but up-to-date for any new relationships. Another shortcoming is private company data, which makes up the majority of the dataset. Private firms are not required to release quarterly financial results, hence there are gaps in our company-level data.

Through each of our tests of the global value chain, we discovered significant methods businesses could use to increase the robustness and resilience of their individual supply chains, which increases the overall global supply chain robustness and resilience. We ranked companies by several centrality measures in Section 2.4.1.2, and in our literature review in Section 2.2.3 we established that company position within the overall supply chain affects

company supply chain robustness. We found that the global supply chain is highly adaptive, indicating high resilience in the supply chain. This finding of high adaptiveness can improve businesses' understanding of the shifting environment in which they operate. Through our analysis of community structure, we developed methods of finding significant communities globally, as well as within specific industry sectors. Our literature review also established that a company's supply chain diversification is tied to its supply chain robustness, hence companies working with partners from a variety of community structures improve their own own supply chain robustness, which can mitigate the effects of disruptive supply chain events.

2.6 Methods

2.6.1 Data Collection. All data included in both networks comes from S&P Capital IQ [17]. Capital IQ provides an Excel interface which allows us to query their database for information. To create the supply chain and strategic alliance networks, we used Capital IQ Excel functions not publically available to retrieve the following data for each company: suppliers, customers, supplier relationships, customer relationships, and strategic alliances. The first step in this process was retrieving all companies stored on Capital IQ's database in order to be able to find all supplier and strategic alliance relationships. We retrieved this data from the Capital IQ website by downloading groups of firms at a time, which was the only manual process required to obtain data. This resulted in more than $12,300,000$ initial firms. Since the Capital IQ Excel Plug-in cannot handle much more than $15,000$ function calls one at a time, we decided to automate the data collection process through Python.

In Python we had to follow several steps. First, we partitioned the data into groups manageable for the Capital IQ Excel Plug-in and created an Excel file formatted properly for the database query. To query the database, we used a DataFrame in the Pandas package [33]. We then opened Excel through Python, loaded the data, ensured that the data saved, and closed the file. At this point, we read the file through Pandas; and verified the data was retrieved correctly, as the Excel Plug-in presented a number of potential errors. If we found

an error, we had Python refresh and open the file again to download the data. We followed this process until the data was formatted as expected. Once this was the case, we formatted the dataset for clarity. If the Excel function was to call suppliers, customers, or strategic alliances, we checked if any companies had come up that were not part of our queue. If we found any they were added to our queue. After this, as an additional error check, we verified suppliers and supplier relationships matched up properly in size, and the same for customers and customer relationships. Any firms that had an unequal number of firms compared to labels were removed from the DataFrame and appended to the queue. This was infrequent but happened occasionally due to Capital IQ updating company data in the time between when we called for the firms in Excel and when we called for the firm labels. Our process ensured we were constantly finding any new firms added to the supply chain or strategic alliance network. After our initial data pull on all firms, we pulled the data for subsequent days starting with the queue from the previous day.

After we retrieved all the data, we then combined the files into MultiDiGraph objects in the NetworkX package [18]. For the supply chain network, we combined the data retrieved for both the supply chain and the customer chain by reversing the direction customer chain, mapping the customer relationships to the corresponding supply chain relationships, and then merging the two networks. This allowed us to use any information found through companies listing a customer that did not list them back as a supplier. The strategic alliance network is undirected and hence was stored as a Graph object in NetworkX. We used NetworkX version 2.7.1 for all file storage.

We used an automated process to obtain data and create graph objects, which allowed us to follow this process daily to get the latest view of the networks for our time series dataset.

The other part of the data collection process was adding company data to the network. This followed a similar process to that of gathering the network. This step can be done with any of the hundreds of Excel functions Capital IQ has created to query their database [17]. For our purposes, we have followed this process with more than 50 of them. We also

automated this process, which only requires an initial input of Capital IQ Excel functions. After this initial input, the Python script partitions the data properly for function calls to Capital IQ in Excel, gathers the data, checks for errors, and stores the data. After the data has been collected, the script then updates all networks found inside a given directory with the new data, using the data as node features.

2.6.2 Computing Statistics and Communities. We then began our data analysis. The global statistics such as those found in Table 2.1, Table 2.2, Figure 2.1, and Figure 2.3 were calculated through the igraph and NetworkX libraries [18, 10]. These packages were also used to calculate the centrality measures seen in Table 2.4 and Table 2.3 as well as the label reciprocities found in Figure 2.4. These methods are fairly straightforward and can be found in the documentation of both packages.

The community detection was done with the leidenalg package, which is built on the igraph package in Python [43, 5]. The leidenalg package uses the Leiden algorithm for community detection based on a given optimization function, which is where the RB method and CPM method differ. The Leiden algortihm for community detection is based off the Louvain algorithm, with improvements to avoid creating weakly connected or disconnected communities and speed improvements.

The Leiden algorithm moves nodes between communities and then tests if the new partition is an improvement over the previous one based on the given optimization function. The RB (Reichardt and Bornholdt's Potts Model) and the CPM (Constant Potts Model) methods are the two optimization method we used in this thesis. For the RB method this optimization function is given by:

$$Q = \sum_{ij} (A_{ij} - \gamma \frac{k_i^{\text{out}} k_j^{\text{in}}}{m}) \delta(\sigma_i, \sigma_j). \tag{2.1}$$

Where A is the graph adjacency matrix, γ the resolution parameter, k_i^{out} is the outdegree of node i, k_j^{in} is the indegree of node j. The $\delta(\sigma_i, \sigma_j)$ function is an indicator function if nodes i and j are part of the same community, resulting in 1 if they are and 0 otherwise.

The RB method is equal to the standard modularity partition method when $\gamma = 1$. Thus, this method is a comparison between our graph and the graph of a configuration null model, and when $\gamma \leq 1$ we put less weight on the null model in the comparison. Hence we are able to create larger and fewer communities than what can be obtained from the standard modularity partition as the resolution parameter decreases, ending with one large community for a connected network if $\gamma = 0$ [36, 29, 47]. Thus for our global community detection 2.4.3.1 by using a resolution parameter of .7 we reduced the amount of weight on the null model to increase community sizes. For the health care community detection 2.4.3.2 we left the resolution parameter at 1, which created more communities with fewer nodes.

The optimization function for CPM is given by:

$$Q = \sum_{ij} (A_{ij} - \gamma)\delta(\sigma_i, \sigma_j). \tag{2.2}$$

Where the $\delta(\sigma_i, \sigma_j)$ function is defined the same way as in the optimization function for the RB method. The goal of this method is to maximize internal edges while keeping small communities. The resolution parameter γ controls this by working as an inner and outer edge density limit. We split a community into two communities 1 and 2 if

$$\frac{e_{1\leftrightarrow 2}}{2n_1 n_2} \leq \gamma. \tag{2.3}$$

Where $e_{1\leftrightarrow 2}$ is the number of edges between communities 1 and 2 and n_1 and n_2 are the number of nodes in each respective community. Thus edge density between communities is less than γ while edge density within communities is greater than γ [42].

Bibliography

[1] D. Acemoglu, V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi, *The network origins of aggregate fluctuations*, Econometrica, 80 (2012), pp. 1977–2016.

[2] V. A. Aggarwal, *Resource congestion in alliance networks: How a firm's partners' partners influence the benefits of collaboration*, Strategic Management Journal, 41 (2020), pp. 627–655.

[3] V. Batagelj and A. Mrvar, *A subquadratic triad census algorithm for large sparse networks with small maximum degree*, Social Networks, 23 (2001), pp. 237–243.

[4] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, *E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials*, Nature Communications, 13 (2022).

[5] V. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *Fast unfolding of communities in large networks*, Journal of Statistical Mechanics Theory and Experiment, 2008 (2008).

[6] Bloomberg, *Global supply chain data — bloomberg professional services*.

[7] K. B. Brian Tjemkes, Pepijn Vos, *Strategic Alliance Management*, vol. 2, Routledge, 09 2017.

[8] A. D. Broido and A. Clauset, *Scale-free networks are rare*, Nature Communications, 10 (2019).

[9] K. Choudhary and B. DeCost, *Atomistic line graph neural network for improved materials property predictions*, npj Computational Materials, 7 (2021).

[10] G. Csardi and T. Nepusz, *The igraph software package for complex network research*, InterJournal, Complex Systems (2005), p. 1695.

[11] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, *Comparing molecules and solids across structural and alchemical space*, Phys. Chem. Chem. Phys., 18 (2016), pp. 13754–13769.

[12] C. Diem, A. Borsos, T. Reisch, J. Kertesz, and S. Thurner, *Quantifying firm-level economic systemic risk from nation-wide supply networks*, Scientific Reports, 12 (2022).

[13] M. Fey and J. E. Lenssen, *Fast graph representation learning with pytorch geometric*, CoRR, abs/1903.02428 (2019).

[14] S. P. Global, *Capital iq database*.

[15] ——, *Panjiva supply chain intelligence*.

[16] ——, *S & p global company services.*

[17] ——, *S& p global market intelligence.*

[18] A. Hagberg, P. Swart, and D. S Chult, *Exploring network structure, dynamics, and function using networkx*, 1 2008.

[19] E. R. Homer, G. L. Hart, C. B. Owens, D. M. Hensley, J. C. Spendlove, and L. H. Serafin, *Examination of computed aluminum grain boundary structures and energies that span the 5d space of crystallographic character*, Acta Materialia, 234 (2022), p. 118006.

[20] T. Hsu, T. Pham, N. Keilbart, S. Weitzner, J. Chapman, P. Xiao, S. Qiu, X. Chen, and B. Wood, *Efficient and interpretable graph network representation for angle-dependent properties applied to optical spectroscopy*, npj Computational Materials, 8 (2022).

[21] F. R. S. Inc, *Factset supply chain relationships.*

[22] H. Inoue and Y. Todo, *Firm-level propagation of shocks through supply-chain networks*, Nature Sustainability, 2 (2019), pp. 1–7.

[23] H. Inoue and Y. Todo, *The propagation of economic impacts through supply chains: The case of a mega-city lockdown to prevent the spread of covid-19*, PLOS ONE, 15 (2020), pp. 1–10.

[24] D. Ivanov, A. Dolgui, B. Sokolov, and M. Ivanova, *Literature review on disruption recovery in the supply chain*, International Journal of Production Research, (2017), pp. 1–17.

[25] W. Kenton, *Strategic alliances: How they work in business, with examples*, Jan 2023.

[26] T. N. Kipf and M. Welling, *Semi-supervised classification with graph convolutional networks*, CoRR, abs/1609.02907 (2016).

[27] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, *The atomic simulation environment—a python library for working with atoms*, Journal of Physics: Condensed Matter, 29 (2017), p. 273002.

[28] Q. V. Le and T. Mikolov, *Distributed representations of sentences and documents*, CoRR, abs/1405.4053 (2014).

[29] E. Leicht and M. Newman, *Community structure in directed networks*, Physical review letters, 100 (2008), p. 118703.

[30] D. Mamedio, C. Rocha, D. Szczepanik, and H. Kato, *Strategic alliances and dynamic capabilities: a systematic review*, Journal of Strategy and Management, 12 (2019).

[31] A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal, *graph2vec: Learning distributed representations of graphs*, CoRR, abs/1707.05005 (2017).

[32] U. of Washington, *Structures of metals*.

[33] T. pandas development team, *pandas-dev/pandas: Pandas*, Feb. 2020.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *Pytorch: An imperative style, high-performance deep learning library*, CoRR, abs/1912.01703 (2019).

[35] V. Randle, *Grain boundary engineering: An overview after 25 years*, Materials Science and Technology, 26 (2010), pp. 253–261.

[36] J. Reichardt and S. Bornholdt, *Bornholdt, s.: Statistical mechanics of community detection. physics review e 74(1), 016110*, Physical review. E, Statistical, nonlinear, and soft matter physics, 74 (2006), p. 016110.

[37] G. Rohrer, *Grain boundary energy anisotropy: A review*, Journal of Materials Science, 46 (2011), pp. 5881–5895.

[38] M. Rojas, A. Routh, J. Sherwood, J. Buckley, and A. Keyal, *Reshoring and "friendshoring" supply chains*, Government trends 2022, (2022), p. 20.

[39] B. Rozemberczki, O. Kiss, and R. Sarkar, *An API oriented open-source python framework for unsupervised learning on graphs*, CoRR, abs/2003.04819 (2020).

[40] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt, *Weisfeiler-lehman graph kernels*, Journal of Machine Learning Research, 12 (2011), pp. 2539–2561.

[41] Y. Todo and H. Inoue, *Geographic diversification of the supply chains of japanese firms*, Asian Economic Policy Review, 16 (2021).

[42] V. Traag, P. Van Dooren, and Y. Nesterov, *Narrow scope for resolution-limit-free community detection*, Physical review. E, Statistical, nonlinear, and soft matter physics, 84 (2011), p. 016114.

[43] V. Traag, L. Waltman, and N. J. van Eck, *From louvain to leiden: guaranteeing well-connected communities*, Scientific Reports, 9 (2019), p. 5233.

[44] D. Watts and S. Strogatz, *Collective dynamics of 'small-world' networks*, Nature Journal, 12 2011.

[45] A. Wieland and C. Durach, *Two perspectives on supply chain resilience*, Journal of Business Logistics, 42 (2021).

[46] W.-Y. Wu, H.-A. Shih, and H.-C. Chan, *The analytic network process for partner selection criteria in strategic alliances*, Expert Syst. Appl., 36 (2009), pp. 4646–4653.

[47] J. Xiang and K. Hu, *Limitation of multi-resolution methods in community detection*, Physica A: Statistical Mechanics and its Applications, 391 (2012), pp. 4995–5003.

[48] Q. Yang, C. M. Scoglio, and D. M. Gruenbacher, *Robustness of supply chain networks against underload cascading failures*, Physica A: Statistical Mechanics and its Applications, 563 (2021), p. 125466.

[49] M. Zopf, *1-wl expressiveness is (almost) all you need*, (2022), pp. 1–8.