EECS4414 Project Progress Report: Analysis of Annual Global Trade through Information Networks

Matt MacEachern
York University
4700 Keele St.
Toronto, ON, M3J 1P3
mcmaceac@my.yorku.ca

Andrew Jaramillo
York University
4700 Keele St.
Toronto, ON, M3J 1P3
andrew2jaramillo@gmail.com

ABSTRACT

(to be added when full results have been obtained from the study)

KEYWORDS

Global Trade, Link Prediction, Network Evolution, Trends, Emerging Markets, Market Community Detection

1 INTRODUCTION AND MOTIVATION

The analysis of global trade has been an essential part of the development and maintenance of economies around the world. Through recessions, war, global catastrophes, sanctions, etc., the global trade network can give interesting and insightful information not only at a historical level, but when looking towards the future of the world economy. This information is valuable, and no doubt being analyzed by economists to determine trends in world trade. This trade network from a high level may not seem as complex and interesting as some networks with higher node counts, however the relationships between nodes are incredibly complex, and even small changes to the topology of the graph could have huge implications when it comes to the overall structure of the network.

This development has interesting implications when looking from the perspective of information networks that trade has developed. At a country by country basis, the trade of goods forms a strongly connected, directed graph with edge weights representing the annual amount of goods sold by one country to another. An example of what this graph may look like after it has been developed throughout this project is shown in Figure 1. The graph itself may not seem grandiose or groundbreaking when graphs of millions of nodes are being analyzed in the social networks that have emerged over the past fifteen years. However, the network itself is still an incredibly interesting one in that it has been evolving for centuries (albeit the data is not readily available or accurate for years before the emergence of computers). Also, the number of interesting graph algorithms that can be applied to this network is not limited by the number of nodes in the network.

In this project, there are many interesting analyses that can be performed on the network itself. There are many properties of this graph that can be analyzed. These include (but are not limited to): link prediction, community detection, time-series analysis, as well as topological analysis to see if the global trade network follows a certain already well-known model.

2 RELATED WORK

There has been a number of researchers who have done extensive analyses of the world trade web. However, Giorgio Fagiolo from Sant'Anna School of Advanced Studies [1,2,6] has dedicated much of his career to the research of economic networks and in particular the world trade web. His papers were a major contributing factor to this paper and gave direction to the properties that were chosen to be analyzed in our data sets. When looking for further properties to analyze in the data sets that we have discovered through the world bank (WITS), Fagiolo's research will be one of the first places that will be considered due to the expertise he has shown through his research.

Lars Backstrom and Jure Leskovec [7] also provided incredibly useful information when considering link prediction and analysis using random walks.

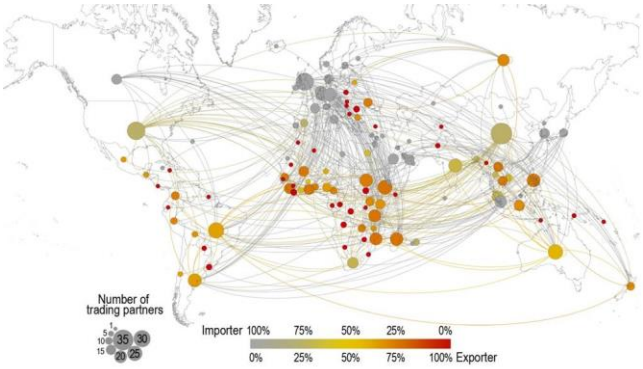

Figure 1: What the global trade network could look like after the development of the graph.

3 METHODOLOGY

3.1.1 Graph Representation

The idea behind representing the data on global trade as an information would be to have each node representing a country. An edge would be between two countries if those countries trade with each other as discovered through the data we collect (see section 4.1 for further information on this data). That is, if node A has an edge connecting to node B, countries A and B are trade

partners in some way. The graph ideally will be directed with edge weights applied to each edge. Node A has a directed edge E to node B with weight W if country A has exported a dollar amount of goods equivalent to W in dollars to country B in that year. Another idea would be to have the net total amount traded between the two countries A and B and have an undirected edge between the two, but this method would be less descriptive.

In this project, there was a total of 195 officially recognized countries that were represented in the graph. Out of those 195 countries, five countries were not represented on the WITS website that was used to retrieve the data for this project (see section 4.1 for details). These countries were: Democratic Republic of Congo, Liechtenstein, Monaco, North Korea, Serbia and Timor-Leste. For purposes of consistency with respect to the data, it was decided not to go to another data source in order to get the export data for these countries, and instead they were excluded from the graph as nodes.

To build the graph there were excel spreadsheet files for each country for a particular year (190 total files, 2014 was chosen to be the year for the initial graph) that were parsed in order to see what countries that particular country was exporting to, and the amount that they were exporting. If the country A had a row in the excel spreadsheet indicating that they were exporting amount W to country B in that particular year, an edge from A to B with weight W was added to the directed graph.

3.1.2 Visual Representation

The visual representation of a graph can be insightful in many cases and can give the viewer an idea of the topology of the graph. Although in many situation when working with graphs with a very large amount of nodes it can be less insightful. In the case of the WTW that was the topic of our study, there were only 190 nodes in our graph, so a visual representation could be useful. In order to obtain a meaningful visual representation of the world trade web, the coordinates of the center of each of the 190 countries was found, and this information was fed into NetworkX.

Figure 5 shows what was obtained when showing the plot of the world trade web in 2014. The nodes in red make it easy to see the outline of each continent and provide a visual of how densely connected each of these nodes are. Due to the high degree of each node on average, the graph is very dense, which makes it difficult to see each connection visually. However, we thought it was an important part of our analysis to add.

3.2 Link Prediction

Building recommendation engines has been studied extensively in different fields. Common methods involved link prediction on heterogeneous graphs. Also, [7] attempts to improve this method, include supervised random walk, where the algorithm assumes the network is homogeneous, and hence the random walk has no constraints. Additionally, link prediction can also be looked through information diffusion. This principle could be applied in trade relationships, where the propagation of the information is the trade itself. Meaning, does the trading patterns of country A affect country B and the countries spanning from B in the trade network.

To experiment with link prediction in the graph, we first start by looking at the common methods: Jaccard's coefficient, Adamic/Adar and Preferential attachment. These three algorithms all require a graph for their input, and their output is a set of three elements (u, v, p). Here, "u" and "v" is the edge in question and p is the score given to the edge. Meaning, what is the likelihood for this edge to occur in the future. For this study, two graphs representing the exports for 2008 and 2014 where used. First, we started by using the 2008 graph as an input to the algorithm and the 2014 graph to test the results of the algorithm. For all three algorithms the p score ranged from 0 to 43255, which is quite large. Additionally, when the results were tested against the 2014 graph, the number of correctly predicted edges was 1038 and the incorrect prediction was 2489. Moreover, the edges that were correctly predicted did not have a similar p value. This meant that some p values for the incorrect prediction was similar or larger than the correct prediction links. Thus, choosing a correct p value for a cut off required additional work. For example, given the 2008 graph's prediction, what is a proper p so that the maximum number of correct links and minimum number of incorrect link can be chosen.

To correctly choose a "p", that maximizes the number of correct link predictions, the following steps where chosen. First, sort the output of the algorithm by increasing p value. Next, construct two arrays named correct_array and incorrect_array. The correct_array is constructed by looping through the output and keeping a count of the correctly predicted edges. If the edge is correctly predicted, then count = count + 1 and correct_array[i] = count; otherwise, if the edge was incorrectly predicted then correct_array[i] = count. The same is done for the construction of incorrect_array, except the loop is from length {len(output) - len(correct_array)} to 1. Additionally, if an incorrect edge is predicted then incorrect_array[i] = npCount - 1, and if edge at i is correctly predicted then incorrect_array[i] = npCount where npCount is the total number of incorrect edges predicted. Finally, loop through both arrays and divide the i value such that correct_array[i]/ incorrect_array[i]. Once this is done, the correct p value to use is the one in the i location where correct_array[i]/ incorrect_array[i] crosses from a less than one to more than one. Refer to Figure 6 for a visual representation.

3.3 Community Detection

Finding communities in any graph is not a simple task, and the WTW network is no different. Although detecting communities using algorithms such as the Girvan–Newman algorithm is very helpful, trying to infer meaning from those findings and knowing when one has sufficiently subdivided a network into communities is not an exact science. Nonetheless, studies have been done specifically on the WTW to find the communities that inevitably exist due to trade deals, geographical location, sanctions, etc. [6]. Barigozzi et al. have performed an interesting analysis on 14 commodities, and the communities that arise due to these commodities in the WTW [6].

In practice, the Girvan-Newman in its simplest form did not provide the most interesting results, with the communities detected being similar to that of a randomly generated graph. Communities of size one would be detected, with a larger central community being the rest of the graph. The single country communities that were being detected were the countries that were not as developed as other countries and were not trading with as many countries as those in the central community. An

example of a country in this situation was South Sudan. Of course, the Girvan-Newman algorithm can be tweaked in order to provide different edge removal criterion, such as heaviest edge or higher edge weights representing stronger ties instead of weaker ties between nodes. However, this did not improve the result and resulted in similar outputs for the communities detected.

This result makes sense logically. Since the WTW is similar in topological properties as that of a randomly generated G(n,p) graph with a higher p value. This randomly generated graph has a high global and local clustering coefficient, just like the WTW, and each node has a high degree. This means that communities detected would not be the most informative as there is no community pattern being developed compared to a more methodically structured network. Keep in mind so far only total export amount has been used to create the graph analyzed in this study. In later iterations of this project, community detection may be performed on the graph formed from the data collected for different commodities in order to see the major communities in areas such as minerals and agriculture.

3.4 Time-Series Analysis

Trade relationships and their evolution over time is an important aspect of the WTW in understanding how the network came to be in its current state. Fagiolo et al. do an incredibly in-depth analysis of the evolution of the WTW from 1981 to 2000 which uncovered numerous interesting facts about network itself [2]. They uncovered that certain many countries have weak trade links, while there seems to be a core structure of rich countries that are more highly connected to other countries in the network. This goes back to the idea of community detection presented in 2.3. For this project, a more recent analysis of the WTW would be interesting in light of recent economic events (specifically the financial crisis of 2008), and how the WTW adapted and evolved in response to these events.

An important model relating to the evolution of the WTW is the fitness network model discussed by Garlaschelli and Loffredo [5]. This model states that each node in the network has an inherent competitive factor called the nodes' fitness. This measure is related to the idea of "the rich get richer" in that nodes that have a higher fitness tend to attract stronger links at the expense of other nodes. The exact math behind this method will be left for when the data is actually collected, however this could potentially be a very interesting factor in looking at the annual network evolution of the WTW.

This section will be expanded on further, but for now our analysis of the world trade web is limited to a single year (2014).

3.5 Topological Analysis

Topology analysis is a quintessential part of any graph analysis, and the topology of the WTW is no different. The structure of the WTW will be verified through the analysis of the data set discussed in 3.1, however many researchers have discovered that some well-known properties of the WTW are that it seems to follow the power-law distribution, has a high clustering coefficient, and follows the small-world network model [1, 3]. Another interesting phenomenon found in these networks in the past has been the correlation between GDP per capita, and the centrality of these nodes in the WTW network [1]. It has been revealed that countries with higher GDP per capita tend to have a more central position in the network and have more trade relationships (edges) in the network than lower GDP countries.

3.5.1 Node Degree Distribution

Most of the nodes in the graph have a high node degree, meaning that they have trade relations with many countries around the world. Since the graph is directed, both the in degree and the out degree of the graph can be analyzed. The out-degree distribution is shown in Figure 2 and shows than many of the countries have many trading partners. There are also many countries that have a relatively low amount of trading partners compared to countries such as Canada and the United States.

The in-degree distribution seems to be much more uniform than the out-degree, with far more nodes having similar in-degrees. This distribution is shown in Figure 3. The disparity between countries when comparing the number of countries they import from is far less than the disparity seen in the out-degree distribution, representing how many countries they are exporting to. This could be due to several factors. Many developing countries may not have the trade relations with many larger, more established first world countries that may have a grip on the market for certain commodities. This could also be due to the fact that many countries have sanctions with certain countries which would prevent them from trading with certain countries.

Surprisingly, the average in-degree and the average out-degree are very similar even though the distributions look different at first glance. The average in-degree (number of countries the node imports from) is 109.7884, while the average out-degree (number countries the node exports to) is 109.78836.

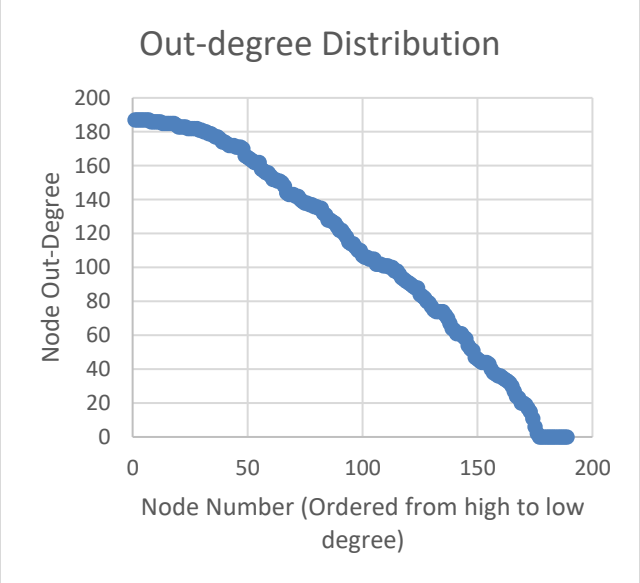

Figure 2: Node out-degree distribution (2014)

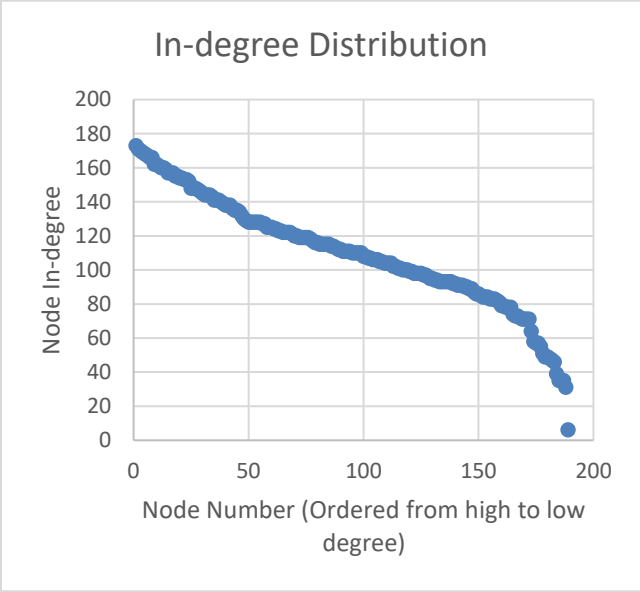

Figure 3: Node in-degree distribution (2014)

3.5.2 Clustering Coefficient Distribution

As expected, the clustering coefficient for most of the nodes was very high, since the graph is very highly connected, and nodes are close to each other in connection. The graph needed to first be converted into an undirected graph in order to obtain the clustering coefficient for each node, but the topological properties relating to the clustering coefficient remained unchanged.

The average clustering coefficient for the world trade web from 2014 is 0.857597497.

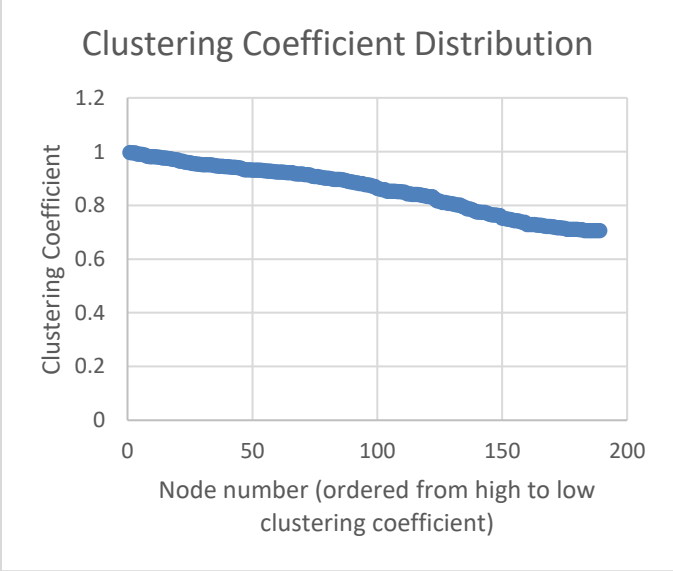

Figure 4: Clustering coefficient distribution (2014)

3.6 Diameter of the Graph

The diameter of the WTW was an interesting point of analysis since it showed how closely connected the world is in the modern era. A question going into this analysis was if a country is not trading with one another, how close are they via their other trading partners and who they are trading with? This measurement was not only important in answering that question but also brought to light some flaws in the data that we had received from the WITS website.

In order to calculate the diameter of a directed graph using NetworkX's diameter function, the entire graph needs to be strongly connected, meaning every node can reach every other node. Unfortunately, just over 93% of the nodes were in the central strongly connected component of the WTW. It was discovered that although countries were reporting that they were trading with a number of countries, those countries themselves didn't have export data on the WITS website. This meant that the out-degree of these nodes were zero (see Figure 2). This caused sinks to form in the graph, meaning that once you reached those nodes, you could not leave. Ultimately it was decided that these nodes should stay in the graph, since countries are trading with them and the data could prove important. However, the diameter function would need to be run only on the central strongly connected component consisting of 177 out of the 190 nodes.

Ultimately, the diameter of the directed version of the graph from 2014 was three, with the undirected version being two. This was an incredibly interesting point of data. We knew previously that the graph was incredibly well connected but knowing that any two countries were only separated by at most three edges put into perspective that although the world is massive, trade has connected nations of the modern world very close together.

4 EVALUATION

4.1 Data Set

The data set uncovered for the purposes of this project comes from the World Bank and is managed by the World Integrated Trade Solution or WITS [4]. WITS allows users to retrieve data on a country by country basis, and filtering on a number of aspects. The data for most countries dates back to 1989, which will suffice for the purpose of our study considering we will be more interested in a recent analysis of the WTW. WITS also allows for filtering on certain product categories such as fuels, chemicals, plastics, etc. to allow for a more in depth analysis if necessary. relationships (edges) in the network than lower GDP countries.

The data from WITS is in excel spreadsheet format, with each country having spreadsheets on their import and export data for a given commodity in a given year. In order to obtain the data in an efficient manner, a script was written in order to query the website for the data from a given year, and the script would download all 190 files needed to gain export data on each individual country. The 190 needed files were determined by cross referencing the ISO3 code for the country and only querying the countries that were listed in a file that contained only the countries we specified (namely, the 190 recognized countries that we knew had data on the WITS website).

5 CONCLUSIONS

The world trade web contains a wealth of aspects and properties that allow it to be analyzed and studied from different perspectives and for different purposes. This study uncovered some of the already well-known properties of the world trade web as well as some lesser known aspects related to the community structure when looking only at the total annual export by country. Time-series analysis results have yet to be determined, but we are confident that we will be able to extract the results from various years in a later iteration of this study. The years following the financial crisis of 2008 will be sure to yield interesting results in terms of the overall structural change and evolution of the WTW, as well as other topological properties such as edge weights.

REFERENCES

[1] Giorgio Fagiolo, Javier Reyes, and Stefano Schiavo. 2009. World-trade web: Topological properties, dynamics, and evolution. Physical Review E 79, 3 (2009). DOI:http://dx.doi.org/10.1103/physreve.79.036115

[2] Giorgio Fagiolo, Javier Reyes, and Stefano Schiavo. 2007. The Evolution of the World Trade Web. Physical Review E 79, 3 (2007).

[3] Ma Ángeles Serrano and Marián Boguñá. 2003. Topology of the world trade web. Physical Review E 68, 1 (November 2003). DOI:http://dx.doi.org/10.1103/physreve.68.015101

[4] "World Integrated Trade Solution (WITS) | Data on Export, Import, Tariff, NTM", Wits.worldbank.org, 2018. [Online]. Available: https://wits.worldbank.org. [Accessed: 07- Feb- 2018].

[5] Diego Garlaschelli and Loffredo Maria. 2004. Structure and evolution of the world trade network. Physica A (2004).

[6] Matteo Barigozzi, Giorgio Fagiolo, and Giuseppe Mangioni. 2011. Identifying the community structure of the international-trade multi-network. Physica A: Statistical Mechanics and its Applications 390, 11 (2011), 2051–2066. DOI:http://dx.doi.org/10.1016/j.physa.2011.02.004

[7] Lars Backstrom and Jure Leskovec. 2011. Supervised random walks. Proceedings of the fourth ACM international conference on Web search and data mining - WSDM 11(2011). DOI:http://dx.doi.org/10.1145/1935826.1935914

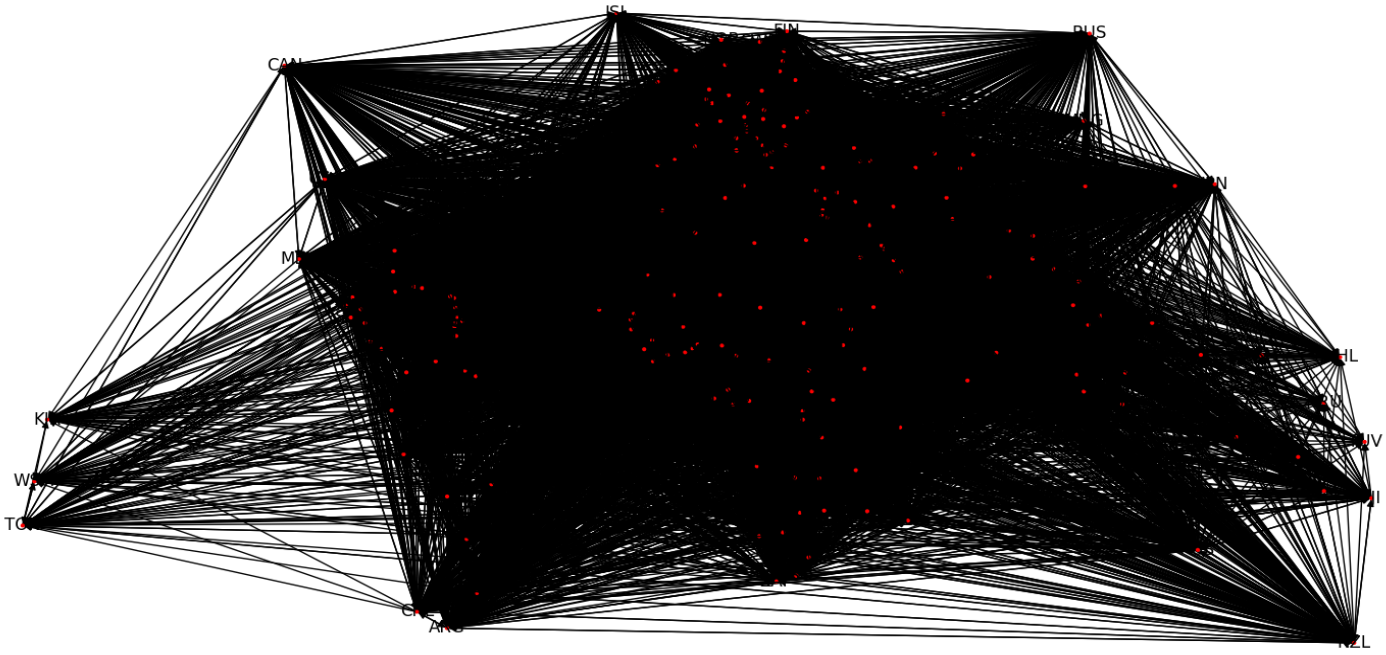

Figure 5: WTW graph extracted from the WITS data set using NetworkX

Output	1	0	0	1	0	0	0	1	0	1	1	0	1
Correct	1	1	1	2	2	2	2	3	3	4	5	5	6
Incorrect rev	7	7	6	5	5	4	3	2	2	1	1	1	0
Correct/ Incorrect	1/7	1/7	1/6	2/5	2/5	2/4	2/3	3/2	3/2	4/1	5/1	5/1	6/0

Figure 6: link prediction example - 1 is correct prediction and 0 is incorrect prediction