An empirical analysis of the FAOSTAT trade network

Darshan Kapashi (darshan@cs.stanford.edu)
Sagar Chordia (sagar.chordia@cs.stanford.edu)

December 10, 2016

1. Abstract

In this paper, we run several experiments on the trade network based on the import and export of forestry goods between countries. This network is unlike other real world networks. There are a small number of nodes ~180 and multiple weighted edges between pairs of nodes. We look at 2 aspects of the network, its topological properties and the community structure in the network. We analyze how these have changed over the years and present an explanation. The trade network is growing over time. The weighted degree distribution follows a power law and we can fit a predictive model on it. We find a positive correlation between node centrality and node weight. We find that the strength of the communities is decreasing over the years. We also find that the world is becoming a smaller place in terms of trade and present evidence for it, in terms of k-cores and hierarchical clustering using strongly connected components.

2. Literature Review

We present a brief review of 2 papers. Section 2.1 talks about an empirical analysis of the virtual water trade network. This is similar to the analysis we have presented in this paper. Section 2.2 talks about an algorithm for community detection based on a random walk. This is one of the algorithms we have applied that was not taught in class.

2.1 Water for food: The global virtual water trade network [1]

This paper studies the network of virtual water trade. The authors use the trade of crops and livestock and the water required to produce those as a proxy for the virtual water trade between countries. They discover that the number of trade connections follows an exponential distribution. There is a power law relationship between the volume of virtual water traded and the number of trade connections of each country. Highly connected nations are preferentially linked to poorly connected nations and exhibit low levels of clustering. They show that a global hierarchy exists such that countries that trade large volumes of water are more likely to link to and cluster with other countries that trade large volumes of water. The paper analyzes the detailed trade matrix published by FAOSTAT and talks about the data collection and transformation process to a graph. We have used the data in a similar way. The major difference in our discussion is that we have conducted a temporal analysis of the network with a focus on the community structure.

2.2 Maps of random walks on complex networks reveal community structure [2]

The paper describes a method to reveal community structure in a graph using random walk as a proxy to succinctly describe information flow. A node group where information flows easily can be grouped in a single module. The algorithm tries to identify modules by finding an optimally compressed description of information flow. The idea is that if we can concisely describe a path, information flow is easier along that path. Each cluster is assigned a unique code and within each cluster, each node gets a unique code. For a random walk within the same cluster, we use a single code and the description is smaller. To find an optimal code, we look for a module partition M of the *n* nodes into *m* modules such that the expected description length of a random walk is minimized. This gives rise to the *map equation* (we direct the reader to [2] for details). To come up with the optimal partition, we use the fraction of time each node is visited by a random walker and using these visit frequencies, we explore the space of possible partitions using a greedy search method. The results are refined using simulated annealing. The authors compare this approach to maximizing the *modularity* metric. It is suggested that when the links represent movement between nodes, the algorithm described here will likely yield better results. When links represent the relationship, the modularity maximization techniques are better.

From here on, we refer to this algorithm as the *infomap* algorithm. This technique fits well with the dataset we are working with, where the edges represent trade between countries. It offers a few parameters such as the markov time which lets us control what kind of communities it will find. The markov time defines how many steps the random walker takes before its position is encoded. Shorter markov times than 1 mean that the average transition rate of a random walker is lower than the encoding rate of its position, such that the same node will be encoded multiple times in a row. As a result, the map equation will favor more and smaller modules. Oppositely, longer Markov times mean not every node on the trajectory will be encoded, and the map equation will favor fewer and larger modules.

3. Description of the data collection process

3.1 The dataset

We have used the dataset of import and export data of forestry goods between countries by the Food and Agriculture Organization of the United Nations [3]. Our focus is on the trade network between countries. The dataset contains information about a few forestry goods for the years 1997-2014. It has data for 182 countries with ~2.42M rows. We make an assumption that the data is complete. This means if we do not see a trade between country A and country B in a given year for some good G, we will assume that the countries did not trade in that year. Each row reports 1 year of import/export of a given good between 2 countries and its quantity. Each row contains the following items:

1. Reporter country: The country which reported this data. There are 182 distinct values.
2. Partner country: The country with which the reported traded. There are 182 distinct values.
3. Item: The good being traded. Examples include Plywood, Newsprint, Sawnwood, etc. There are 14 distinct goods.
4. Element: The quantity of trade. This is given either in terms of import quantity (in tonnes or m3), import value (in USD), export quantity (in tonnes or m3) or export value in USD. This sums to 6 choices in each row.
5. Year: The year of trade. This goes from 1997 to 2004.

3.2 Data transformation

We only consider the export data, because there are inconsistencies in the data, for example, when A reports export to B, it is expected that B will report import with A, but the data does not always reflect this. We take the USD amount as the edge weight. Finally, the edge weights are very skewed. For instance p25 percentile of weight is $35, p50 percentile is $150, p80 is $3000, p95 percentile is $300000. If we do linear weighting of edges then only top 5% of edges contribute 99% weight of graph. This is expected given the economy differences in countries. Some of the algorithms we use don't handle such a skew well. We take the logarithm of the value as the weight of each edge. The edge weights now range from 1 to 20.

3.3 Graph construction

We create 1 graph for each year of data we have. We use 3 different constructions for our analysis:

A. Each country is a node. We add an edge between 2 countries if they have any trade in that year. The graph is unweighted. The edge direction is from exporter to importer country.
B. The skeleton is the same as A. Each edge now has a weight equal to the sum of the log values of the trade amounts.
C. The skeleton is the same as A. If the edge weight as described in B between country X and Y is k, we add k edges between X and Y. For example, if the edge weight between USA and China is 10, we add 10 edges between USA and China.

Graph A is unweighted and directed. Graph B is weighted and directed. Graph C is unweighted, directed with multi-edges.

4. Analysis and results

We apply various techniques we learnt in class and studied in literature review and present our findings in this section. This section is roughly organized as follows.

4.1 Basic properties

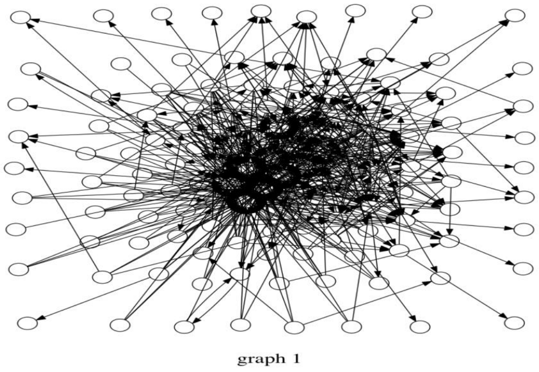


Figure 4.1.1 (a): Visual of the network in 2014

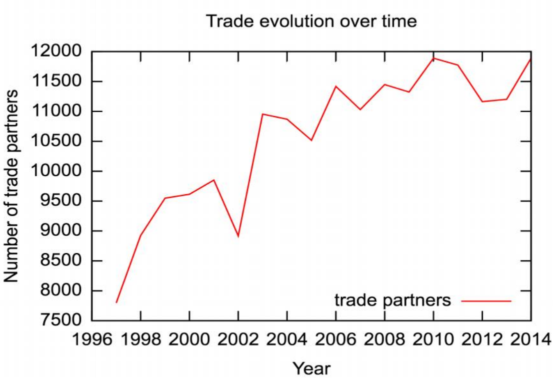


Figure 4.1.2 (a)

4.1.1 Visualization

Figure 4.1.1 (a) is a visualization of the trade network in 2014. We find that a few countries contribute to bulk of international trade. The top countries with the maximum trade, such as USA, China, India, Germany, France, UK and Italy have edges with almost all countries. Whereas there are some small countries (the islands) which mostly import from a couple of neighbouring countries.

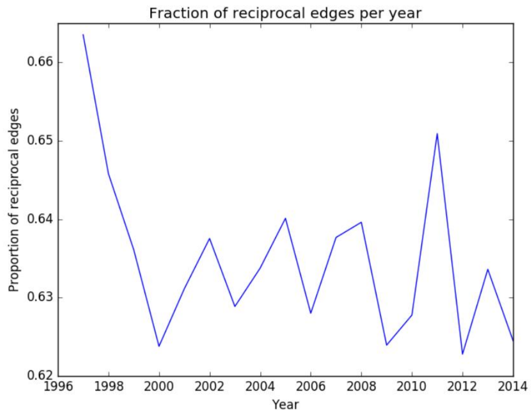
4.1.2 Trade network has increased over time

Figure 4.1.2 (a) plots the number of edges (y axis) in the network over time (x axis). The number of edges has increased over the years, indicative of globalization. We see dip in 2002 which could be explained by the recession around the dotcom crash.

4.1.3 Diameter of the network

The diameter of the network varies between the values of 3 and 4 over the years. This conforms with small world phenomenon [4] where any two countries can be reached in less than 4 hops. The value 3 or 4 is determined by existence of edges from small countries (e.g. Myanmar) to bigger hubs (e.g. India).

4.1.4 Reciprocal edges



If (a, b) is an edge, then (b, a) is its reciprocal edge. We find that over the years, the proportion of reciprocal edges is roughly the same, ranging from 63% to 67%. This means approximately 65% of the edges in the graph of each year have the reverse edge too. For a given good, we generally expect to find a single edge from A to B, since if A is exporting good X to B, we don't expect it to import the same good. However, we have multiple goods and so it is natural that A exports some to B and imports some others. There are the countries in the periphery, see figure 4.1.1 (a), which only import and contribute to the 35% of non-reciprocal edges.

4.2 A predictive model on the degree distribution

4.2.1 Does the degree distribution follow a power law?

We analyze the degree distribution for different possible graphs:

1. Unweighted, directed graph, similar to graph A from section 3.3, except that we only consider a single good in a given year. The degree distribution resembles the power-law distribution. However, now we have too many individual graphs (~250) and this is not a scalable way to analyze the data. See figure 4.2.1 (a) below.
2. Unweighted, directed graph as in graph A from 3.3. It considers all the goods together in a year. This graph doesn't show any power-law distribution. See figure 4.2.1 (b) below.
3. Weighted, directed graph as in graph B from 3.3. The weighted degree distribution resembles a power-law distribution, see figure 4.2.1 (c) below.

These figures plot the number of nodes with the given degree (y axis) vs the node degree (x axis).

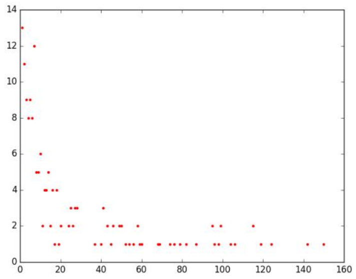


Fig 4.2 (a): Degree distribution for (1)

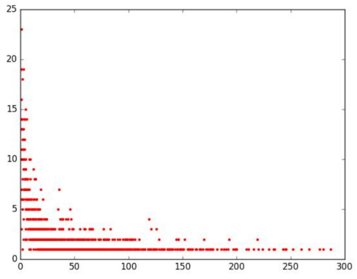


Fig 4.2 (b): Degree distribution for (2)

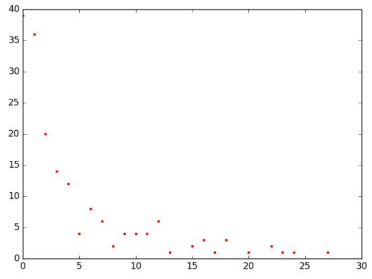


Fig 4.2 (c): Degree distribution for (3)

4.2.2 Can we fit a model to the degree distribution?

We try to fit a model to the degree distribution of this network. If we plot the power law graph on a log-log scale, it will be a straight line. We use this principle and try to fit a line on the log of data. The 3 graphs 4.2.2 (a), (b) and (c) show the fitted line for 3 different years, 1998, 2004 and 2012. The y axis has the number of nodes and the x axis has the node degree. We find that the data fits nicely along the line.
Based on the 18 graphs corresponding to 18 different years, the average power law   is -0.96. We see that the value of increases over time. We fit a linear equation to the   value based on the data from 1998-2010. Using this, we predict the value of   for the next 4 years. For the estimated value of , we report the root mean square error (RMSE), i.e. RMSE on the eval data, and compare it with the RMSE for the training data from first 10 years. As seen in the table below, our estimation of is very good.

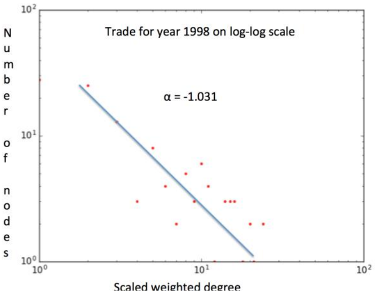


Figure 4.2.2 (a)

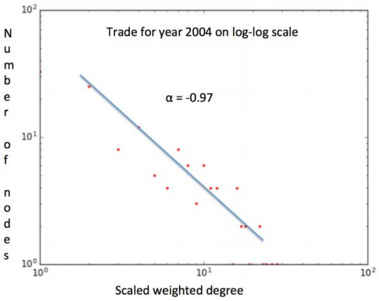


Figure 4.2.2 (b)

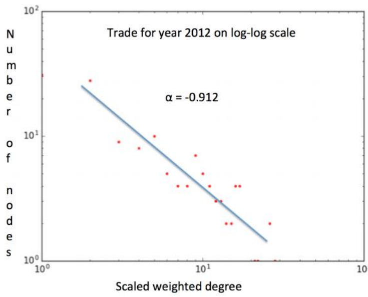


Figure 4.2.2 (c)

Data	Year		RMSE (error)
Training	1997 - 2010	-1.086 to -0.921 (increasing with year)	2.34
Evaluation	2011 - 2014	-0.913 to -0.883 (increases with year)	2.41

4.3 Betweenness centrality

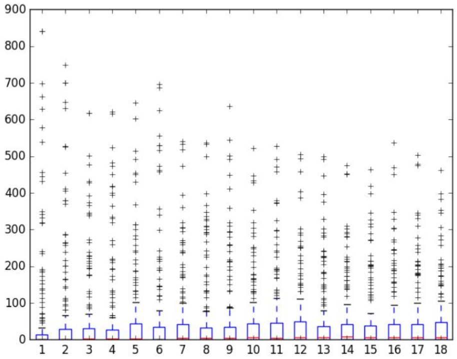


Figure 4.3.1 (a): Box plot of node betweenness centrality vs year

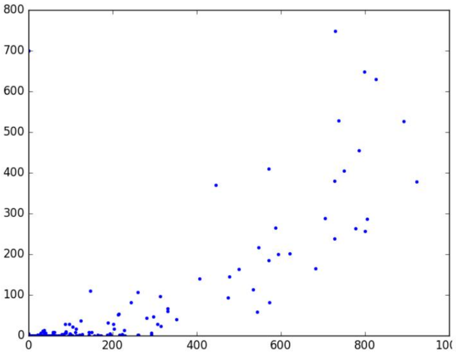


Figure 4.3.1 (b): Node centrality vs node weight

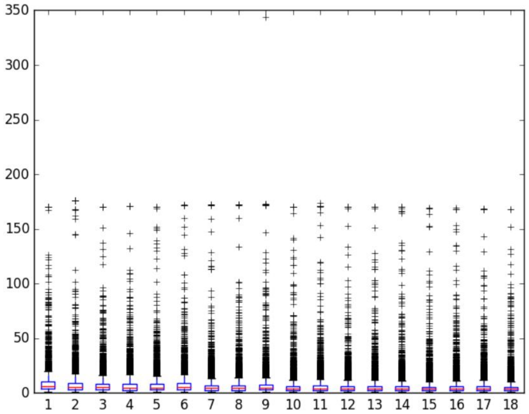
4.3.1 Node betweenness centrality

The data clubs a few countries under the name Other and Total FAO. We find that a few nodes have a very high node centrality, however the p75 value is less than 50. The top countries by node centrality are *Other*, *Total FAO*, *UK, Germany, France, Italy, China* and *USA.* In figure 4.3.1 (a), the x-axis has years from 1997 (year 1) to 2014 (year 18). For each year, we have a box, the lower bound is the p25, the red line is the mean and the upper bound is the p75. The points above p75 are the more extreme values. We see that the maximum value of node centrality decreases over time. Over the years, new edges get added as more countries start trading with each other and hence the shortest paths through a node decrease over time.

Let's see if node centrality is correlated with the node weight. We define the node weight as the sum of the weights of all outgoing edges of a node. Figure 4.3.1 (b) shows the node centrality on the y axis versus the node weight on the x axis. It has a positive correlation of 0.78 (in 1998). Across the years, the correlation coefficient varies between 0.75 to 0.83, although there is no trend. There is indeed a strong correlation.

4.3.2 Edge betweenness centrality

Edge centrality shows a similar distribution as node centrality. As shown in the figure on the right (box plot of edge centrality vs year number), most of the edges have low centrality. The p75 value is less than 20 paths. However, a fraction of the edges have a very value and this skews the distribution. We find that edges from small countries (such as Tonga, Samoa and Chad) to big countries (such as UK, Italy, etc.) are the ones with the high centrality. This shows that an edge connecting a small country to a very big country becomes very important for all shortest paths from those small countries. We also see high edge centrality for edges between countries with lot of trade e.g. USA and China. As a result even though we see correlation between edge centrality and weight of edge, it's much weaker, compared to the correlation we saw in the previous section 4.3.1.



Another interesting observation is that in year 2005, the maximum edge centrality is exceptionally high with a value of 343.7, whereas in other years the maximum is less than 200. This is the edge between French Polynesia to China. This edge exists only in this year which might correspond to the trade agreement between the 2 countries in 2004 [5]. China is a hub and many shortest paths between two countries in the world go via China. Some small countries such as Tonga and Samoa have an edge to French Polynesia. Their path to China is shorter now because of this new link. If this link doesn't exist their shortest paths to other countries are still from China but using an indirect route. This suggests that a link to a big hub could become important in such situations.

4.4 Neighbour overlap

For a given edge, the jaccard similarity of the set of neighbours of the source node and the destination nodes is defined as the edge overlap. Intuitively, edge overlap means how closely connected are two nodes of an edge (we also refer to it as the edge strength). Edge overlap has a value between 0 and 1. As in figure 4.4 (a), the p50 of edge overlap is between 0.2 and 0.3. The box plot also shows that the p25, p50 and p75 of the edge overlap increases over time. This is in line with globalization and the growth of the trade network 4.1.2. This is a stronger result than 4.1.2 since it also considers structural properties of the graph. We evaluate the hypothesis; edge strength is correlated with edge weight. In [6], the authors compare the telephone network of USA and observe that the edge weight (number of calls between two nodes) varies linearly with the edge strength. We compare our edge weight (USD value of trade between two countries) to edge strength. We observe a positive correlation of 0.27 when we take logarithm of the weights. If we take the absolute values, the correlation is 0.13. The hypothesis holds true on our graph but the correlation is not as strong as in [6]. Figure 4.4 (b) plots the edge strength (y axis) versus the logarithm of the weight (x axis). A higher weight has a higher edge strength, however there is no clear linear relation.

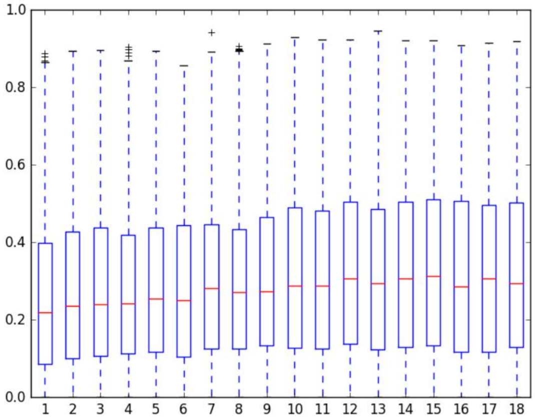


Figure 4.4 (a): Box plot of edge overlap vs year number

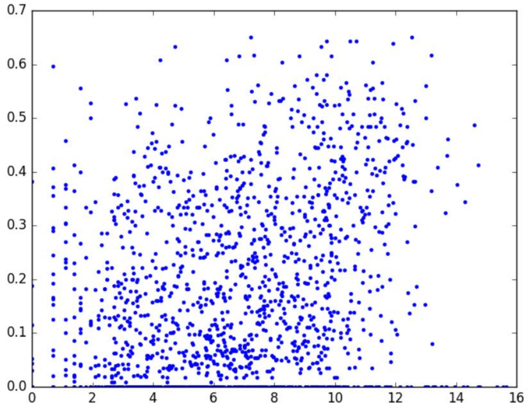


Figure 4.4 (b): Edge overlap vs edge weight

4.5 Community Structure

We use 3 approaches to analyze the community structure of the network: Strongly connected components (SCC), K-cores in the graph and the Infomap method [2].

4.5.1 Communities using SCCs

A strongly connected component (SCC) is defined as a set of nodes such that there exists a path from every node to every other node. Across all the years, we see that there is a single big SCC with ~170 countries. The remaining ~10 countries are not a part of any SCC. The size of the largest SCC varies from 168 to 170.

Now, we start removing edges with the highest edge centrality one by one and build hierarchical clusters by identifying the SCCs. As we remove edges, we continue seeing a similar SCC structure as described above, i.e. 1 giant SCC whose number of nodes keeps decreasing and the isolated nodes. Initially, we find that a few small countries such as Bhutan and island countries such as Samoa, French Polynesia, etc. get separated from the big SCC. In the next step, bigger but *not so well off in terms of trade* countries such as Afghanistan separate out. In the remaining steps, medium to big sized countries start getting excluded from the big SCC, but the structure persists, 1 giant SCC and 1 node SCCs. After several steps, there are 2 SCCs seen, 1 with big countries such as USA and China and the other with countries from Africa and Middle Eastern Asia.

4.5.2 K-cores

A k-core of a graph G is a maximal connected subgraph of G in which all vertices have degree at least k. It is one of the connected components of the subgraph of G formed by repeatedly deleting all vertices of degree less than k. We analyze 2 graph structures. For both of them, we have 2 plots. The first one plots the number of nodes in the k-core (y axis) versus k (x axis). The second one plots the highest k for which the k-core was found.

Figures 4.5.2 (a) and (b) show the trend for the unweighted, directed graph as in 3.3 graph A. Figures 4.5.2 (c) and (d) show the trend for the unweighted, directed graph with multi-edges as in 3.3 graph C. In both cases, we see that there are more nodes for higher values of k and there is a clear trend there, figure 4.5.2 (a) and (c). (we show 5 out of 18 years of data, the rest is similar). Also, in both the graph structures, we see that the biggest k-core grows by more than 25% since 1997, figures 4.5.2 (b) and (d). This is in line with the increasing trade over the years (4.1.2) as well as the small world phenomenon (4.1.3).

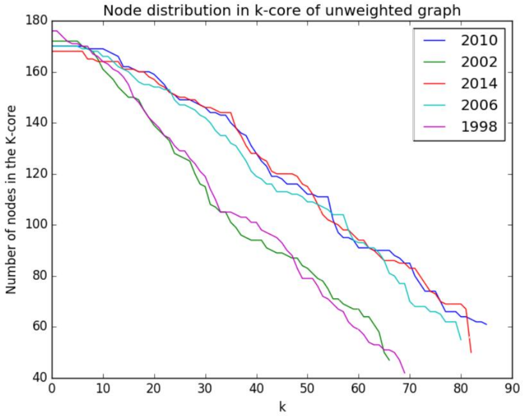


Figure 4.5.2 (a)

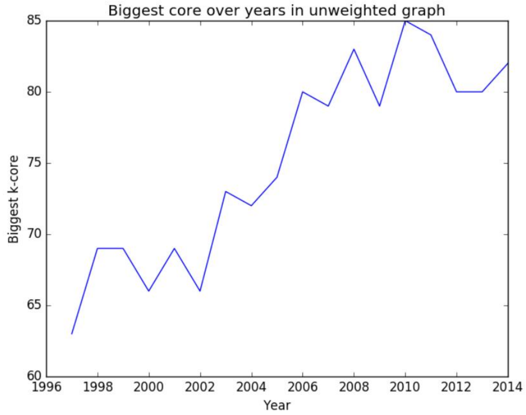


Figure 4.5.2 (b)

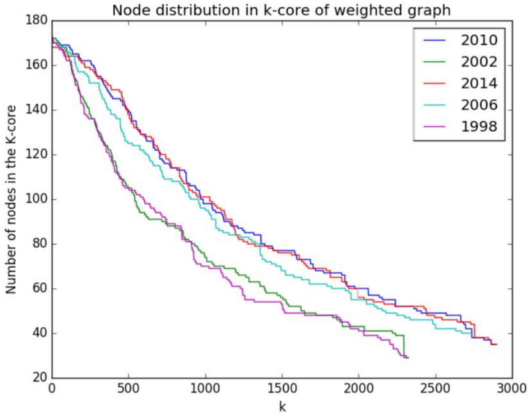


Figure 4.5.2 (c)

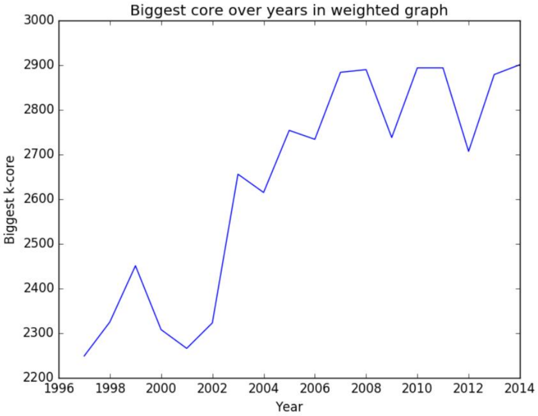


Figure 4.5.2 (d)

4.5.3 Infomap method

Based on the comparative analysis presented in [7], the authors suggest that the *Infomap* method by Rosvall and Bergstrom [2] is the best performing. We apply the infomap method on our dataset. The algorithm takes in a parameter called the markov-time. With a markov time parameter of 1, the random walker is more likely to visit farther parts of the graph. The larger countries of the world trade over long distances too, which manifests as a single giant community. When we lower the markov time parameter, we count the same node multiple times. The data represents trade of forestry goods such as wood, so countries are generally connected to nearby countries, and we weigh these connections higher. This results in communities of countries which are geographically close by.

On 2014 data, with a markov time of 1 and 10 communities, we find that 146 countries are in a single community. The rest of the communities have 6, 5, 4, 2, 1, 1, 1, 1 and 1 countries in it. Using a markov time to 0.92, the distribution shows more interesting patterns with 65, 27, 25, 13, 11, 7, 6, 6, 5 and 3 countries. The countries in these communities are geographically close, for example, the biggest is composed of countries from Europe. The community with 27 is South-East Asia plus Australia and the one with 25 members is North and South America.

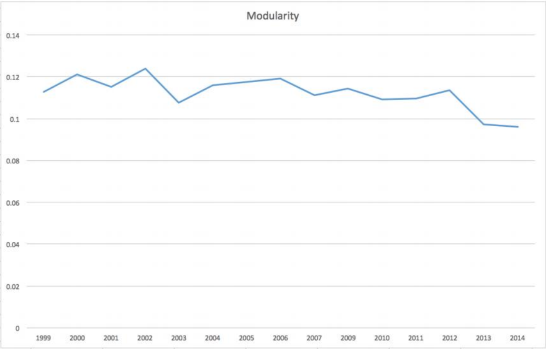


Figure 4.5.3 (a)

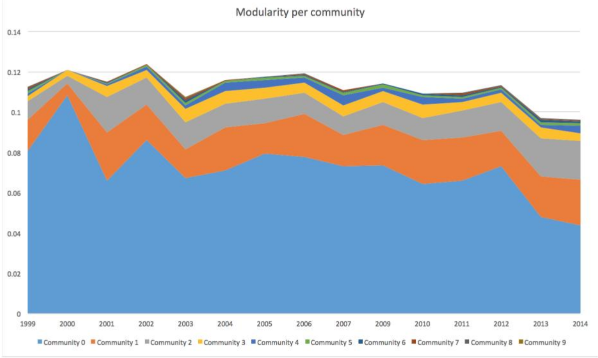
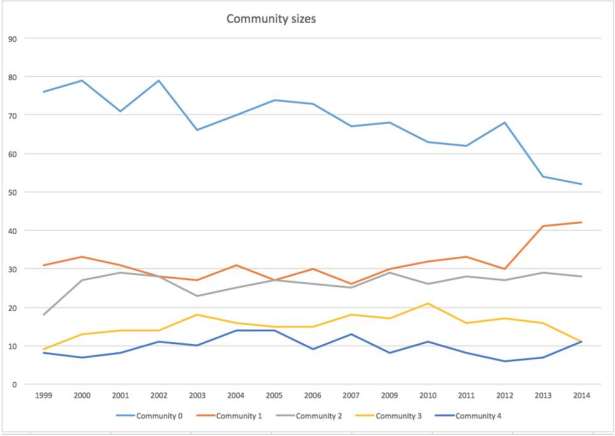


Figure 4.5.3 (b)

Applying this to the data in 2004, we find the markov parameter of 0.92 gives a community structure with 1 giant community, similar to 2014 with markov parameter 1. This means countries were trading less with their immediate neighbours and more with countries they favored, even though they were far. Lowering the markov parameter to 0.85, we see a similar membership distribution as in 0.92 in 2014.

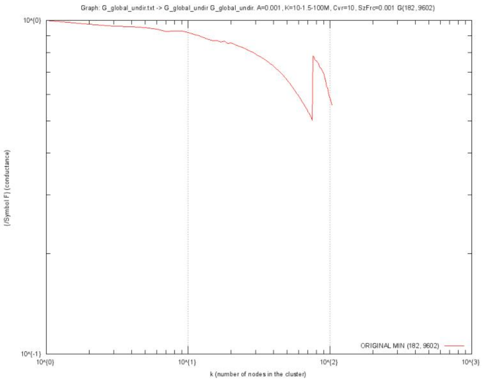
Modularity [8] is a measure of the strength of a division of the network into modules. A higher modularity relates to closely connected communities. Figure 4.5.3 (a) shows the modularity of the communities after applying community detection to every year. The modularity is going down over time. Looking at the modularity per community in figure 4.5.3 (b), we alse see that the modularity of the biggest community is also decreasing over time. The second and third biggest communities are growing stronger by themselves. We can also look at the size of the communities over the years and we see a similar trend, where the biggest community is shrinking and the second biggest community is getting bigger. As more countries trade with each other, countries will have ties with other countries outside their community too. This makes the modularity of their own community smaller which explains the decreasing trend. In addition, as countries form stronger bonds with each other, we have more hopes of finding communities, i.e. we had 1 big community in 1997 and in 2014, we have 2 large communities.



4.6 Overlapping communities

We use k-clique community to identify neighboring communities in this network. We start with k = 3 and keep increase the value of k up to 30. We observe that for lower values of k we see 1 big community comprising of ~170 nodes. As k increases we still see only one big community with decreasing size. As k reaches around 25 we still see 1 community with about 30 countries in that cluster. As k reaches 30 we don’t see any community at all. We plot the network community plot (NCP) to check if conductance drops with increase in k. Conductance is defined as number of edges going out of the cluster divided by the total degree of the nodes in the cluster.

$$\phi(S)=\frac{|\{(i,j)\in E; i\in S, j\notin S\}|}{\sum_{s\in S} d_s}$$



The figure on the right plots the conductance (y axis) versus k (x axis). The graph is assumed to be undirected. Our actual graph is directed, so this is not an accurate representation, however it gives us a good sense about the graph structure. The conductance decreases with increase in k, reaches a minimum and then increases drastically again. The conductance is minimum at k = 80. This implies a very closely interconnected cluster of 80 countries exist. But as soon as one node from this cluster is removed conductance increases dramatically. Upon inspection we find that this cluster of 80 countries is comprised of all medium and big economies. The island countries and economically smaller countries are not a part of this cluster.

5. Conclusion

We analyzed the trade network of forestry goods based on the data reported by FAOSTAT. We discovered a few interesting properties in the network. We observe the small world phenomenon in terms of the diameter of the network. The degree distribution on the weighted graph follows a power law and we were able to fit a predictive model with a small error. Nodes with a higher weight play a more central role in the network, similarly edges with a higher weight have a higher centrality too. As countries trade more, there is a large strongly connected component and communities are nested rather than overlapping. We also find that the community strength decreased and the largest community became smaller. As future research direction, we would develop a model on this trade network, fit it with the given data and see if we can predict the numbers and properties for future years.

6. References

[1] Konar, M., et al. "Water for food: The global virtual water trade network." *Water Resources Research* 47.5 (2011).

[2] Rosvall, Martin, and Carl T. Bergstrom. "Maps of random walks on complex networks reveal community structure." *Proceedings of the National Academy of Sciences* 105.4 (2008): 1118-1123.

[3] http://faostat3.fao.org/download/F/FT/E

[4] Watts, Duncan J., and Steven H. Strogatz. "Collective dynamics of ‘small-world’networks." *nature* 393.6684 (1998): 440-442.

[5] http://dfat.gov.au/geo/french-polynesia/pages/french-polynesia-country-brief.aspx

[6] Onnela, J-P., et al. "Structure and tie strengths in mobile communication networks." *Proceedings of the National Academy of Sciences* 104.18 (2007): 7332-7336.

[7] Lancichinetti, Andrea, and Santo Fortunato. "Community detection algorithms: a comparative analysis." *Physical review E* 80.5 (2009): 056117.

[8] Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), 026113.

[9] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

[10] Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society.*Nature*, *435*(7043), 814-818.

Individual contributions: We believe we have each contributed equally to the project, in terms of coding, analysis and the writeup.