Networks and Measured Variables Explain One Another

James Kennedy
Washington, DC

Abstract: An approach is described for using measured variables to explain network structure, or vice versa. The approach considers whether entities tend to be linked to entities that are near them in the space of the measured variable or variables.

Keywords: networks, statistics

I. INTRODUCTION

Living organisms in nature tend to attempt to organize themselves optimally. Social behavior, from the slime mold to the human being, satisfies complex environmental and biological constraints and supports reproductive success, whether the organisms comprising the population are aware of their own motivations or not.

Human social behavior is especially difficult to understand, for the obvious reason that we, as human beings ourselves, are biologically predisposed to particular kinds of explanations of our own behavior. We may objectively study a society of social insects, for instance, experiment with them and identify the parameters of their swarming, foraging, nesting, and so on, but when it comes to human society the best we can do is observe, experiment, and philosophize in as objective a manner as possible.

Annotation of the structure of human organizations and human/computer networks can potentially shed light on important characteristics of our social motives, even when phenomena produce no conscious impression on participants in a system. But structure alone is not enough to explain itself. When we have measured features of the node-entities of the network, we should be able to use that knowledge to explain how they came to be arranged in the observed way – or the network structure may shed light on the variability of the observed features.

The present paper proposes an approach to the use of measured variables to explain the pattern of a network. The example is a network of international trade between countries, but the method is applicable to any network where something quantifiable is known about the networked entities.

II. NETWORKS

A network in the present context is a representation of the relationships among a set of objects. Links can be

directed (asymmetric) or bidirectional (symmetric matrix), binary or numeric, signed or not. The present study develops a network of reciprocal trade partners using binary (0,1) bidirectional links and then calculates numeric weights for them.

In a network, distance is measured in terms of the number of steps it takes to get from one node to another. The positioning of the nodes in a display is considered to be arbitrary – Figure 1 shows several depictions of the same network. The spatial configuration of the nodes can help a researcher identify important aspects of the network structure, but does not usually mean anything in itself.

FIGURE 1. THE ARRANGEMENT OF NODES IN A NETWORK DIAGRAM IS ARBITRARY.

A network can be represented as a square adjacency matrix, where ones and zeroes indicate whether nodes are linked or not, as seen in Table 1. This table is the matrix representation of the network shown in Figure 1, with node indexes included for clarity.

TABLE 1. ADJACENCY MATRIX REPRESENTATION OF NETWORK IN FIGURE 1, WITH INDEXES.

	1	2	3	4	5	6
1	.	1	0	0	0	0
2	1	.	1	1	0	0
3	0	1	.	0	0	0
4	0	1	0	.	1	0
5	0	0	0	1	.	1
6	0	0	0	0	1	.

The present study uses a kind of international trade network to demonstrate a research method. Because trade is a system of reciprocal exchange, viewed at various levels of aggregation – nodes can be individuals, firms, industries, countries, products, ideologies, etc. – it is a good topic for network representation. Though it is not mainstream economics, there are a number of researchers using network representations of international trade [2, 3, 4, 7].

The present network was developed using 2012 Comtrade commodities export data at the HS2 level, including all countries, all partners [5]. Data were collapsed to a matrix of total exports for each reporter $\times$ partner. Where exports to the other were greater than 0.2 percent of both countries' total exports, they were considered "trade partners" and a link was represented between them; the goal was simply to produce a matrix that was not too dense or sparse. Of the 293 countries in the database, 69 countries had trade partnerships that qualified. These relationships were represented as a diagonally symmetrical binary adjacency matrix, graphed in Figure 2.

Note that the present definition of trade partnerships tends to penalize very large and very small countries; while a large country such as China might make up a significant part of a small country's exports, the converse is not so likely, and so the reciprocal relationship is not established. Turkey, a medium-sized trading country, had the most reciprocal partners, at forty-seven.

III. SCATTERPLOTS

In a scatterplot, the values of variables are their coordinates in a Cartesian system (variables may be scaled). The standard scatterplot is usually two-dimensional but three dimensions can be depicted using linear perspective techniques or an interactive computer display. In one dimension, a scatterplot is just a number line. Regardless of dimensionality, a vector of values equals a point in a Euclidean space.

FIGURE 2. GRAPHICAL REPRESENTATION OF RECIPROCAL INTERNATIONAL TRADE NETWORK.

The utility of a scatterplot for the present approach is that the distance between vectors can be easily calculated when considered as the Euclidean distance between points in a conceptual scatterplot. While the scatterplot is usually called upon to discover correlation, it can also show clustering of measured entities in the variable space, and other kinds of patterns.

FIGURE 3. AN EXAMPLE OF A SCATTERPLOT.

IV. EUCLIDEAN NETWORK

If the entities represented by nodes of a network have some measured numeric values, the network can be represented as a connected scatterplot. An example is shown in Figure 4. The same scatterplot of measured values is represented, embedded in two different connection networks. The lower figure uses curved lines to un-overlap the edges and make the connections visible.

FIGURE 4. NETWORKS PLOTTED IN SCATTERPLOTS.

In both examples, note that the edges connecting the nodes on the left instance are shorter than the edges on the right. This means that node entities are connected to partners that are closer to them in the variable space: they have more similar scores on the measured variable.

In a social network, two forces can explain why neighbors would cluster together in a variable space:

- Homophily: people form ties with people who are similar to themselves, and
- Norm formation: people become more similar to those they interact with

The result of these forces is a correlation between distance in the Euclidean space and distance in the topological connection space. The measurable effect in a Euclidean network is that the total (or mean) Euclidean path length gets shorter.

Euclidean distance is calculated as the length of a hypotenuse in the Pythagorean Theorem:

$$Dist_{i,j} = \sqrt{\sum_{d=1}^D (x_{id} - x_{jd})^2}$$

where x_{id} is the measured value for entity i on dimension d . This distance can be calculated in any number of dimensions. In one dimension, as in the present example, it

is equivalent to the absolute value of the difference between two values.

V. A MEASURED VARIABLE: THE HUMAN CAPITAL INDEX

In 2013 the World Economic Forum published some measures of the countries of the world [6], rating them on

- Education,
- Health & Wellness,
- Workforce & Employment, and
- Enabling Environment.

The four measures were aggregated into a single score called the Human Capital Index. This index assigns each country one evaluative score. The coordinate system is a one-dimensional number line. The question then is this: does this measure give us any understanding of the structure of the network of trade partnerships?

VI. COMPARISON STANDARD #1: SIMPLE RANDOM NETWORKS

Countries were conceptualized as lying on a number line where their position was their HCI rating. The distances between all pairs of connected countries were calculated and divided by the number of connections, with the finding that the empirical mean path length = 0.1621 HCI units.

Having determined that short average Euclidean distances between nodes suggests a relationship between the network structure and the variables defining the nodes' position in the space, the question comes up, how do we determine what is "short?" It is apparent that shortness is defined in comparison to some standard. One statistical way to ask the question is, what is the probability that this observed value is due to chance?

An obvious test is to compare the observed Euclidean path length to the distribution of some random networks' path lengths. Toward that end, 100 networks were generated, using random permutations of the subdiagonal bitstring, reflected into the upper diagonal. Because the upper and lower halves of the matrix are symmetrical, the lower half of the adjacency matrix was recast as a single-dimensional binary vector, and random permutations were taken. The matrix was then filled with the lower half comprising the permuted bitstring and the upper half being the reflection of that. The original HCI vector was then used to position the nodes, and the mean Euclidean path length was computed for the network.

The distribution of the measure is shown in Figure 5. The observed path length in the empirical data is approximately seven and a half standard deviations below the mean of the random populations.

FIGURE 5. THE DISTRIBUTION OF MEAN PATH LENGTHS FOR 100 NETWORKS GENERATED USING A UNIFORM RANDOM DISTRIBUTION, COMPARED TO THE EMPIRICALLY OBSERVED NETWORK'S MEAN PATH LENGTH.

VII. COMPARISON STANDARD #2: RICH-GET-RICHER NETWORKS

Barabási [1] has recommended use of a “rich-get-richer” algorithm to generate random networks. In an iterative process, nodes that have more neighbors are more likely to gain new neighbors. This has the tendency to result in scale-free networks having qualities of networks found in the real world. Thus, an algorithm was programmed where a node was selected at random, then ten candidate neighbors were evaluated; the one with the most neighbors was assigned to connect to the target node. The resulting distribution was seen to resemble a power-law distribution for degree.

As shown in Figure 6, the empirical mean path length was 3.3373 standard deviations below the mean of the random population.

FIGURE 6. THE DISTRIBUTION OF MEAN PATH LENGTHS FOR 100 NETWORKS GENERATED USING A RICH-GET-RICHER RANDOM DISTRIBUTION, COMPARED TO THE EMPIRICALLY OBSERVED NETWORK'S MEAN PATH LENGTH.

VII. ALTERNATIVE QUESTION: RANDOM VECTOR OF MEASURES

So far the research question has been, how likely is the observed network, given a particular vector of HCI values? But we can also ask: how likely is the observed vector of

HCI measures, given the empirically observed trade network? This question can be addressed by taking random permutations of the HCI vector, and testing their fit in the observed network. So now Albania may have Russia's HCI score, Vietnam will have Canada's, and so on, and countries will be connected as was observed in the reciprocal trade network.

One hundred permutations of the HCI measure were assigned to the trade network, and mean path lengths were calculated, as before. As shown in Figure 7, the observed path length was 3.94 standard deviations below the mean of the random population.

FIGURE 7. THE DISTRIBUTION OF MEAN PATH LENGTHS FOR 100 RANDOM HCI VECTORS, COMPARED TO THE EMPIRICALLY OBSERVED NETWORK'S MEAN PATH LENGTH.

Using a Z test, all of the results presented so far are significant at the $p < 0.0001$ level. In fact, in every comparison the empirically observed mean path length is lower than the lowest random case.

VIII. WILCOXON MANN-WHITNEY TEST

The Wilcoxon Mann-Whitney (WMW) test is a nonparametric statistical technique for comparing two samples where the distributions cannot be assumed to be normal; this test is robust even where the sample sizes are very different. In the present case we can ask the research question in a slightly different way: given the empirical network and measured vector, are the connected pairs of nodes significantly closer together than unconnected pairs?

In the current network of 69 countries there are 2,346 pairs of countries, where 658 pairs are connected in the network and 1688 are not. These were arranged into a dataset of distances, and the connected pairs were compared to the not-connected pairs. As seen in Table 2, the linked pairs were a significantly shorter distance apart than the pairs that were not linked. The WMW test found $Z = -7.5912$, $p < 0.0001$ (actually $3.1694E-14$). As seen in Table 2, the distances between connected countries were significantly smaller than the HCI distances between countries that were not reciprocal trade partners.

TABLE 2. DESCRIPTIVE STATISTICS FOR COMPARISONS OF DISTANCES OF HUMAN CAPITAL INDEX VALUES BETWEEN PAIRS OF CONNECTED AND NOT-CONNECTED COUNTRIES IN THE TRADE NETWORK.

	N	Mean	Median	Stderr
Linked	658	0.5865	0.4925	0.0173
Not-linked	1688	0.7753	0.6780	0.0131

IX. CONCLUSIONS

We can safely conclude that the relationship between the observed trade network and the Human Capital Index in the real world did not arise due to chance. Trade partnerships form between countries that are similar to one another on the Human Capital Index, and perhaps countries that trade become more similar in the HCI variable. It may not be a hard phenomenon to explain narratively – countries trade with other countries that are developmentally similar to themselves – but the economic hypothesis is not the focus of the present analysis.

The causal relationship is signified by the fact that node-entities that are connected to one another are significantly closer to one another on an explanatory measured variable, or, to put it the other way (because this is a correlational relationship), node-entities that are similar in the measured variable are more likely to be connected to one another.

A couple of little-used statistical tests, namely Geary’s C and Moran’s I , are used to assess “spatial autocorrelation,” a concept similar to the present one, where the proximity of some objects correlates with their connectedness. The present approach using the WMW test has the advantage of being easy to understand, easy to implement, and it answers the actual question we want to answer: are connected objects more similar to one another than would be expected if connections were made by chance?

This preliminary study used a one-dimensional distance, but the multivariate version of the method has been seen to work well and provides sometimes surprising insights. Note for instance that in one dimension, e.g., latitude, Seattle and Budapest are within a few miles of one another, but in two dimensions they are more than 5,000 miles apart. Similarly, analysis of one variable may miss or veil a relationship that becomes evident in the multivariate case, or may suggest a relationship where none exists.

Historically, students of networks and of statistics have spoken in different terms, and knowledge developed in the two fields has largely not diffused across their mutual disciplinary boundary. The present approach integrates the two perspectives on data analysis.

Knowledge of correlation between network structure and some measured variables may be pragmatically applied by arranging networks, for instance organizing teams of workers, in a way that minimizes the mean path length of the aggregate in terms of some criterion variable. Or, the other

way, such knowledge could suggest the need for training and other measures to bring coworkers in line with their colleagues and improve group performance.

In the particle swarm, as in human society, we observe that individuals come to cluster together with their topological neighbors in the problem space, and that the size of their search steps decreases as they cluster. This clustering tendency can be exploited in an optimization algorithm, and as the current discussion suggests, it can also be adapted for data analysis and the understanding of causal linkages between structure and value.

In the context of blended human-computer systems, where humans and machines communicate through some kind of network, it may be possible to identify the qualities of machines and humans that make them understandable to one another and facilitate productive communications. When appropriate variables are identified then shortest-path-length networks can be set up to propitiate the achievement of results.

REFERENCES

- [1] Barabási, A-L., and Réka, A. (1999). Emergence of scaling in random networks, *Science*, 286, 509-512.
- [2] De Benedictis, L., and Tajoli, L. (2011). The World Trade Network. *The World Economy*, 34(8): 1417-1454.
- [3] Dueñas, M., and Fagiolo, G. (2013). Modeling the International-Trade Network: a gravity approach. *Journal of Economic Interaction and Coordination*, 8, 155-178.
- [4] Rauch, J.E. (2001). Business and Social Networks in International Trade. *Journal of Economic Literature*, 39, 1177-1203
- [5] United Nations Statistical Division. (2013). Data retrieved October 21, 2013, from <http://comtrade.un.org/>
- [6] World Economic Forum (2013). Human Capital Index. <http://reports.weforum.org/human-capital-index-2013>
- [7] Zhu, Z., Cerina, F., Chessa, A., Caldarelli, G., Riccaboni, M. (2014). The rise of China in the international trade network: a community core detection approach. *PloS one* 9 (8), e105496.