Network Analytics: Analyzing the Stock Market data

Anita Mezzetti

The goal of this task is to learn the causal structure among different technology companies by observing their stock price over a period of time.

First of all, we import the data. Unlike the last assignment, where our data were in the form *(node, sample, time)*, in this case the shape of our data is `(7, 190, 12)`, where 7 is the number of market days (T), 12 the number of companies and 190 the number of samples for each company for each time. Therefore, in order to use the algorithm of the last assignment, we resample our data to obtain a shape equal to `(12, 190, 7)`. We would have got the same result if we would have left the data as they were and we would have modified the algorithm instead. We try both ways and we prove that the result does not change, as could be expected.

We continue the description considering the first case, in which we resample our data. To find the causal structure, we need to calculate the directed information between each pair of nodes (each node is a company). In case all processes are jointly normal, as in this problem, we are given that the information between X and Y given Z is

$$I(X \to Y||Z) = \sum_{1}^{T} \log \frac{|\Sigma_{y_1^t z_1^{t-1}}||\Sigma_{x_1^{t-1} y_1^{t-1} z_1^{t-1}}|}{|\Sigma_{y_1^{t-1} z_1^{t-1}}||\Sigma_{x_1^{t-1} y_1^{t} z_1^{t-1}}|}; \tag{1}$$

where Z includes all the other companies. Then, Z's size is 10.
We also know that $\Sigma_{y_1^t z_1^{t-1}}$ is the covariance matrix of $(y(1), ..., y(t), z(1), ..., z(t-1)$ and that $|\Sigma_{y_1^t z_1^{t-1}}|$ is its determinant.

If we analyse the first time, we get that the vectors which arrive to $t-1$, as z_1^{t-1}, are empty. Then for t=1, the addend is

$$\log \frac{|\Sigma_{y_1^1}|}{|\Sigma_{y_1^1}|} = 0.$$

For this reason, we start our sum from $T = 2$ (which in python is 1 considering we start from 0) to avoid the `Out of Bounds Exception`. Now, we describe the functions present in the code:

0.1 Functions `det_cov`, `log_part` and `direct_information`

First of all, we should remember that Python starts indexing from 0. For this reason in the sigmas of Equation 1, indexes start from 0.
The function `det_cov` finds the norm of the Covariance Matrix of a combination of $x_0^{x_{t0}}$, $y_0^{y_{t0}}$, $z_0^{z_{t0}}$,

which can be used also for y_0^{yt0}, z_0^{zt0} if we do not pass any parameter regarding x. This method selects the values only concerning the variables and the times we are interest in and save them a variable called *data*. We have that each line of data (composed of 190 values) is one of the variables $(x(0), ...x(x_{t0}), y(0), ..., y(y_{t0}), z(0), ..., z(z_{t0})$. The function returns the determinant of the covariance matrix of *data*. In other words, the method `det_cov` gives back $|\Sigma_{x_0^{xt0} y_0^{yt0} z_0^{zt0}}|$, considering the python indexing.

The method `log_part` is in charge of calling the previous function four times to find the four Σ in Equation 1, merge them together and do the log in order to create each addend of the sum. So, `log_part` calculates, for a certain time t, $\log \frac{|\Sigma_{y_0^t z_0^{t-1}}||\Sigma_{x_0^{t-1} y_0^{t-1} z_0^{t-1}}|}{|\Sigma_{y_0^{t-1} z_0^{t-1}}||\Sigma_{x_0^{t-1} y_0^t z_0^{t-1}}|}$.

The function `direct_information` add all the addends together, starting from $t = 1$ for the reason mentioned above, and gives back

$$\sum_{t=1}^{T} \log \frac{|\Sigma_{y_0^t z_0^{t-1}}||\Sigma_{x_0^{t-1} y_0^{t-1} z_0^{t-1}}|}{|\Sigma_{y_0^{t-1} z_0^{t-1}}||\Sigma_{x_0^{t-1} y_0^t z_0^{t-1}}|}$$

Therefore, `direct_information` returns the directed information between x and y, which are always two nodes, given the set of all the other nodes z.

It is noteworthy that we have decided to include in z all the other companies. As a matter of fact, considering more information , which means including more companies, leads to an higher information. If we include completely uncorrelated nodes in z, they would have zero impact. The idea is: adding more information will always give a zero or positive impact on the correlation. So, we can directly include all the other companies and not try every combination among them.

After the definition of the previous functions, we compute the direction information of each pair of companies. If this information is above a certain threshold, we can consider this information as positive, otherwise it is zero. The directed information can never be negative by construction. The information matrix is represented in Table 1.

The rule to construct an edge between two companies x, y is:

$$\forall x, y/ \quad x \neq y \quad \wedge \quad z = \text{nodes} - x, y \qquad \text{If} \quad I(x \to y|z) > \text{threshold} \quad \Rightarrow \quad \texttt{add\_edge(x,y)}.$$

Obviously, we obtain different graphs using different threshold.
For instance, setting `threshols = 0.45` we obtain the graph showed in Figure 1.

	1	2	3	4	5	6	7	8	9	10	11	12
1	0.000	0.282	0.250	0.290	0.466	0.499	0.637	0.678	0.552	0.383	0.380	0.175
2	0.337	0.000	0.251	0.428	0.319	0.255	0.087	0.291	0.243	0.191	0.174	0.159
3	0.212	0.384	0.000	0.161	0.571	0.143	0.234	0.400	0.249	0.436	0.345	0.276
4	0.200	0.181	0.318	0.000	0.277	0.159	0.238	0.309	0.420	0.252	0.207	0.277
5	0.344	0.274	0.266	0.336	0.000	0.265	0.317	0.299	0.337	0.399	0.365	0.247
6	0.350	0.415	0.200	0.340	0.300	0.000	0.197	0.635	0.224	0.425	0.248	0.327
7	0.148	0.337	0.313	0.323	0.132	0.298	0.000	0.179	0.198	0.214	0.350	0.245
8	0.480	0.548	0.340	0.360	0.542	0.210	0.368	0.000	0.205	0.284	0.382	0.408
9	0.436	0.342	0.250	0.100	0.319	0.131	0.318	0.199	0.000	0.268	0.273	0.257
10	0.290	0.412	0.473	0.477	0.303	0.270	0.217	0.423	0.319	0.000	0.184	0.212
11	0.230	0.260	0.152	0.164	0.244	0.242	0.136	0.121	0.087	0.169	0.000	0.184
12	0.111	0.145	0.230	0.240	0.107	0.166	0.171	0.155	0.237	0.246	0.292	0.000

Table 1: Directed information matrix. Each value indicate $I(x \to y|z)$ where x is a row and y is a column. For instance: $I(1 \to 2||2,3,4,5,6,7,8,9,10,11,12) = 0.282$

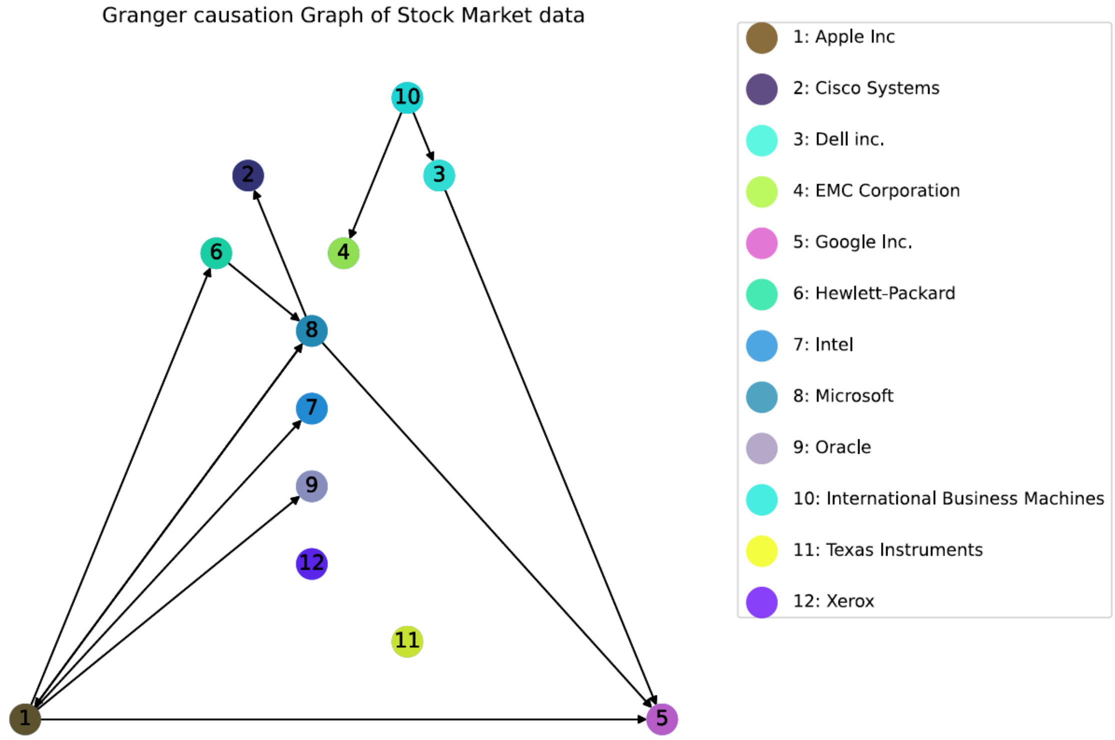

Figure 1: Granger causation Graph of Stock Market data for threshold=0.45.