Network Science - Interdisciplinary Project: Group 7.1 Noodle-Spaghetti

Due on 28^{th} January 2020

Prof. Tomaso Erseghe, Prof. Leonardo Badia, Prof.ssa Caterina Suitner

Elena Camuffo 1234370, Laura Crosara 1234372, Matteo Moro 1234368

Contents

1 Introduction: the dataset

1.1 Background

As **Pasta** has recently become one of the most spread foods all around the world, it makes us investigate how other cultures are adapting the original Italian recipes in order to fit each population desires.
The aim of our group is to analyze the ingredients used for pasta in *three different countries*: **Italy**, **Taiwan** and **Japan**, in order to give an answer to the following questions: which are the *most popular ingredients* used for pasta in those different cultures? Are the ingredients of these cultures *similar or different*?

1.2 Data Collection

To retrieve the data (*ingredients* and *recipes*) we chose three websites:

Italy: `www.giallozafferano.it`,
Taiwan: `www.icook.tw`,
Japan: `www.cookpad.jp`.

In each of these we searched for the **keywords** corresponent to the italian name *pasta* ("pasta","意大利面", "パスタ" respectively for the IT, TW, JP website) and then we took the *first thousand* recipes filtering them by **popularity**. A further refinement has been made because, as far as Italian recipes are concerned, the keyword "pasta" has several uses from appetizers to desserts.

Once all the ingredients of all the recipes have been extracted, it was decided to **break down each type of pasta** (for example: spaghetti, penne rigate etc.) **into its main ingredients** in order to have more precise datasets.

It should also be noted that the data scraping was performed using *Python 3.7* principally using the *BeautifulSoup* module for the manipulation of HTML files of the relative webpages.

1.3 Matrices building

Once collected data we built *two adjacency matrices* for each country:

- **A bipartite matrix** (fig. 1) where the *ingredients* are on one side and the *recipes* on the other side. An ingredient is linked to a recipe if it owns to it.
- **A projection matrix** (fig. 3) where *ingredients* are labels for both columns and rows. The ingredients are linked each other if they own to the *same recipe*. So the links are *weighted*.

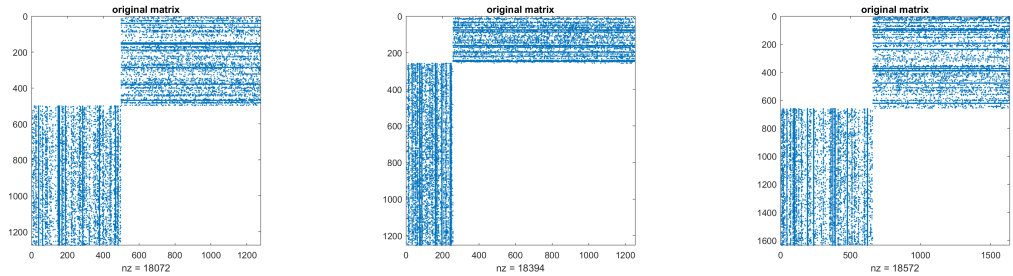
1.4 The Networks

1.4.1 Bipartite

The bipartite networks matrices are shown in figure 1. For Italy and Taiwan the number of analyzed recipes is around 800 and we have a considerable number of ingredients, while as regards Japan, the recipes are even more, but the ingredients dataset is poorer.

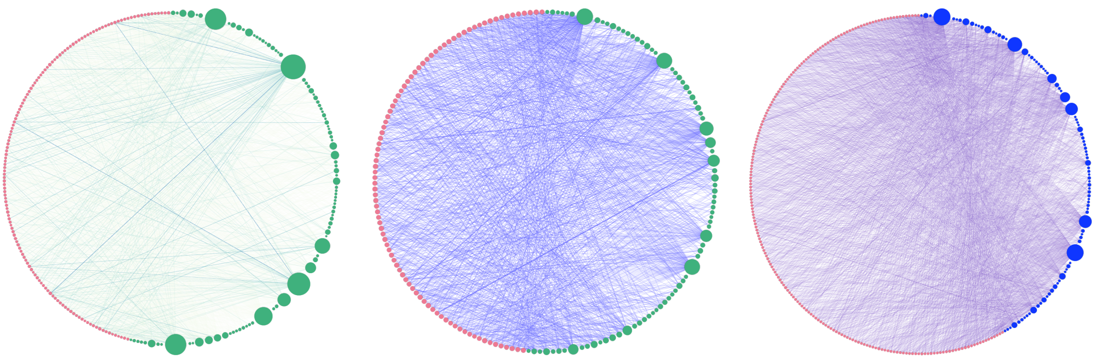
1.4.2 Projection

Figure 4 shows the three network graphs, which are **undirected and weighted**.
We can observe that:



(a) Italy (b) Japan (c) Taiwan

Figure 1: Bipartite Matrices.



(a) Italy (b) Japan (c) Taiwan

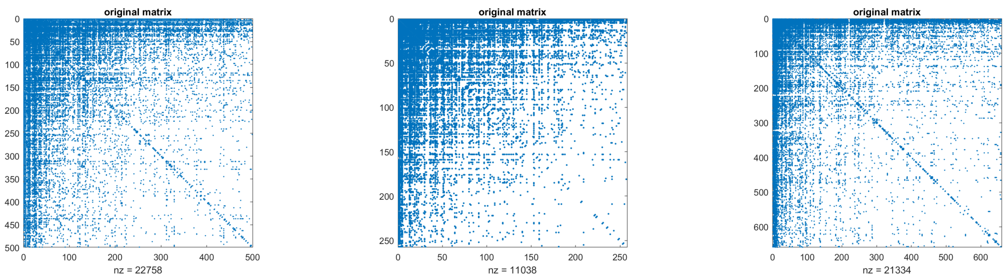
Figure 2: Bipartite network graphs.

- **The Italian** network (fig. 4a) presents a high number of nodes with different degrees and many *links with low weights*.
- **the Taiwanese** network (fig. 4c) presents a *non-connected component* in correspondence with a non-conventional recipe of "pasta" made with *white chocolate*, *Ferrero Rocher chocolate*, *vanilla ice cream* and *strawberry jam*.
- **the Japanese** network (fig. 4b) presents a small number of nodes but also a *huge amount of hubs*.

Table 1 resumes the network parameters.

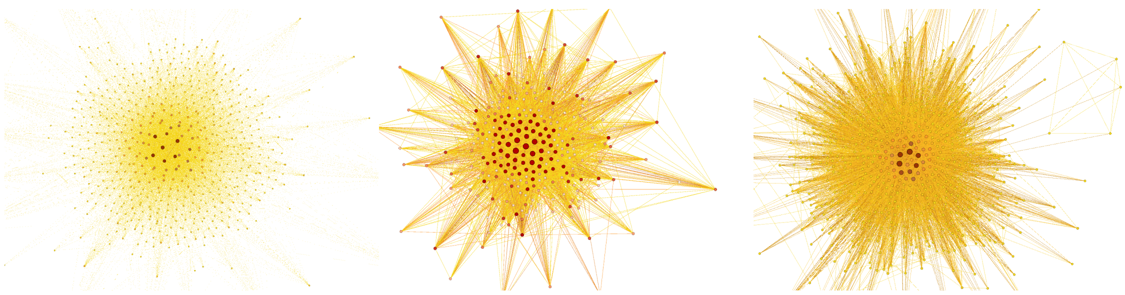
	Italy	**Taiwan**	**Japan**
Number of nodes N	500	659	257
Number of links L	22790	21334	11038

Table 1: Projection Network parameters.



(a) Italy (b) Japan (c) Taiwan

Figure 3: Projection Matrices.



(a) Italy (b) Japan (c) Taiwan

Figure 4: Graph plot of Projection network. Dark edges correspond to more weighted links. Dark and big nodes correspond to high degree ingredients.

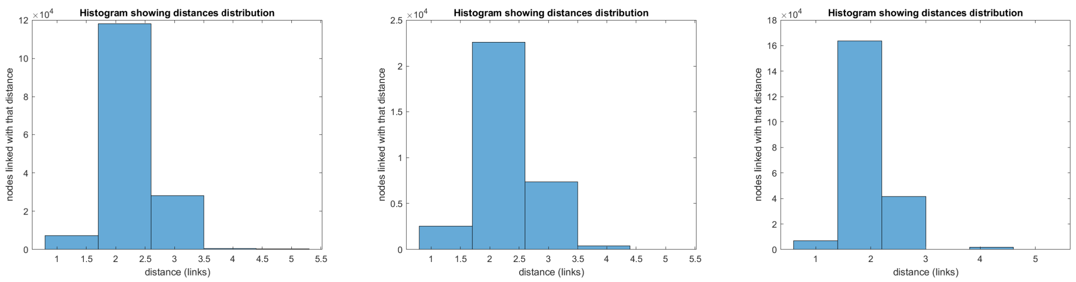
2 First Part

In this chapter we are going to perform a first analysis of the network in general, focusing on the network model and on its general features.

2.1 Diameter and distance distribution

Histogram of figure 5 shows the distance distribution in the three projection matrices and table 2 reports the *diameter* and *average distance* for each network.
We can notice that Taiwan network has diameter ∞ because of its disconnected component but its *giant component's diameter* is comparable with other countries ones.
We can summarize results about distances (fig. 6) as follows:

- For **Italy** the farthest ingredients are *rice water*, *red cabbage*, *ginger*, *pecorino di fossa cheese* and *Sbrinz cheese*.
- Instead, for **Japan** the farthest ingredients are *delicious dore*, *chiza*, *propagule*, *okonomiyaki souce* and *spam*.
  Notice that while in Italy ginger is one of the most uncommon ingredients, it is one of the most common in Japan (it is the 37^{th} most popular ingredient).
- For **Taiwan** the farthest ingredients are the ones related to the *non connected component*: *white chocolate*, *Ferrero Rocher chocolate*, *vanilla ice cream* and *strawberry jam*.



(a) Italy (b) Japan (c) Taiwan

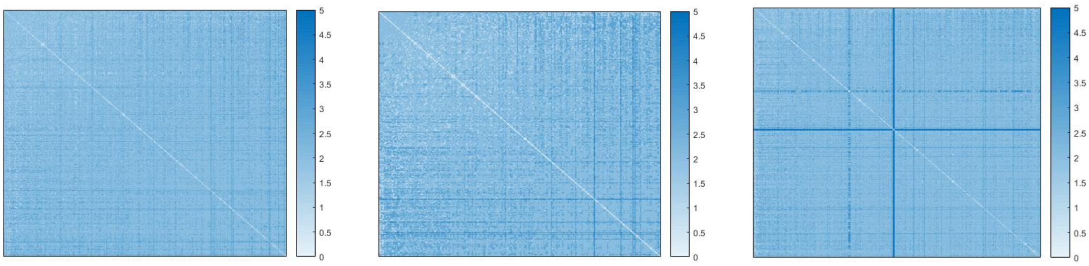
Figure 5: Distance Distribution histogram. Each network has mean distance ≈ 2 according to the analytic results (table 2).

	Italy	Taiwan	Japan
diameter	5	∞ (5)	5
average distance	2.1261	∞ (2,1778)	2.1625

Table 2: Diameter and average distance.

2.2 Network model

Table 3 summarizes the projection network's parameters.
We can see that the *Japanese network holds the highest average degree*, i.e. most of its nodes have a high degree, and consequently its power-law exponent γ *is the lowest one.* Also Italy and Taiwan networks have $\gamma \leq 2$.

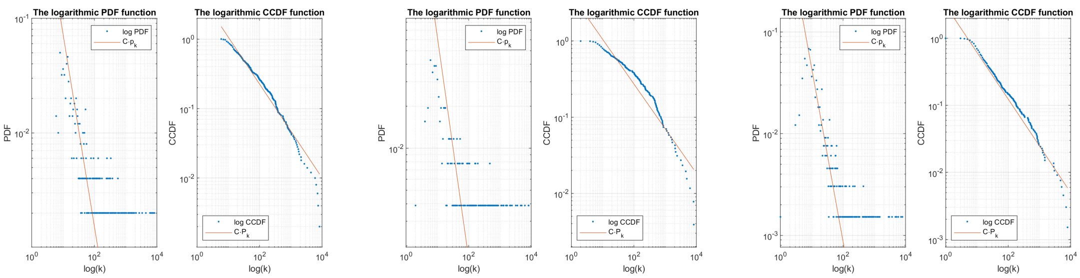


(a) Italy (b) Japan (c) Taiwan

Figure 6: Distance Matrices. *Darker* shades of blue correspond to more distant nodes.

So the networks do not hold properly a scale-free behavior ($2 \leq \gamma \leq 3$), and this is confirmed by the *divergence of the moments* (they increase very quickly as the order grows). Also the variance is very large.

However the networks behaviours seems to follow a power-law quite faithfully (fig. 7).
The three degree distribution functions are *heavy-tailed*, highlighting the presence of hubs. In fact, in such distributions there are lots of nodes with small degree and a few nodes with a very high degree.
The presence of more high degree nodes in the Japanese network can be seen in the logarithmic Complementary cumulative density function (CCDF) plot, decreasing more slowly than the CCDF plots of the other two networks.



(a) Italy (b) Japan (c) Taiwan

Figure 7: Probability density function (PDF) and Complementary cumulative density function (CCDF).

	Italy	**Taiwan**	**Japan**
Average degree $\langle k \rangle$	221.46	138.1487	343.3074
Second Moment $\langle k^2 \rangle$	499021.031	6101.6767	9836.1427
Third Moment $\langle k^3 \rangle$	16790209142.1845	23780605.6996	28773374.6426
Variance σ^2	630603.6884	374737.4802	991670.2129
k_{max}	8249	7879	8194
k_{min}	6	1	2
γ	1.6759	1.7334	1.7059
γ_{sat}	1.7693	2.0503	1.6803

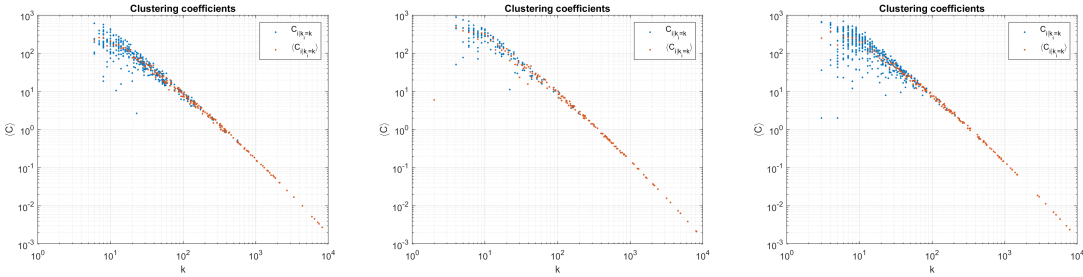
Table 3: Other network parameters.

2.3 Clustering Coefficients

Clustering coefficients were studied both on the projection and the bipartite networks.

On the **projection** they were defined as *the probability that two incident edges are completed by a third one to form a triangle*, i.e. two edges are incident when they end up in the same node.
So, chosen one node, we take two of its neighbours and check if there exist an edge that connects them.

Figure 8 shows a comparison between clustering coefficients and nodes degrees.
We can see that the *clustering coefficient is inversely proportional to nodes degrees*, consequently very common ingredients are likely to have neighbours of different types.



(a) Italy (b) Japan (c) Taiwan

Figure 8: Clustering coefficients. Blue dots are nodes of the network, represented according to their degrees (X axis, in logarithmic scale) and their local clustering coefficient (Y axis, in logarithmic scale). Orange dots show the average clustering coefficient taken over all nodes of the same degree.

	Italy	Taiwan	Japan
Average Clustering Coefficient $\langle C_{i\|k_i=k} \rangle$	78.978	- (129.431)	111.849

Table 4: Average clustering coefficients. In brackets the one referred the to the Taiwan's giant component.

On the **bipartite** network, the clustering coefficients were defined as *the number of existing squares* (made by a node, two of its neighbours and a common neighbour between them) over the total amount of possible squares with that number of nodes.

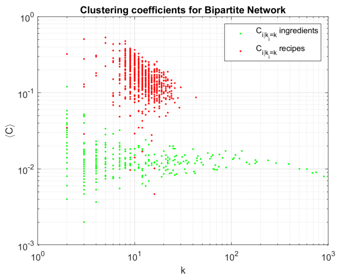
Figure 9 illustrate that result. We can see that *Red dots, related to recipes*, delineate a *degression trend* in all the three networks, i.e., with the raising of the number of ingredients a recipe is linked with, it is less likely that two of these ingredients will be used both in another recipe.
The clustering coefficient *decreases more rapidly in Italy and Taiwan* than it does in Japan.

2.4 Robustness

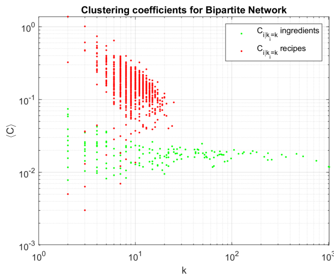
Robustness was studied either for bipartite and projection networks. Particularly we take into account:

- Robustness to **random failures**;
- Robustness to **attacks** (remving hubs first).

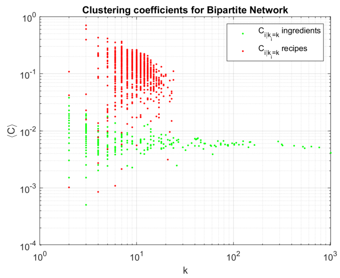
As regards the **projection** network, the results are shown in figure 10.
We can notice that all the networks share a *very similar robustness to random failures.* The result is expected because all the networks are approximatively scale free (few nodes keep the network connected) and



(a) Italy



(b) Japan

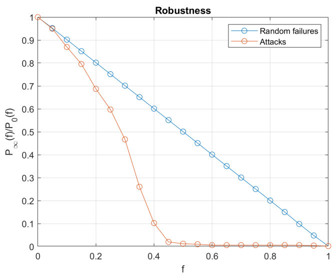


(c) Taiwan

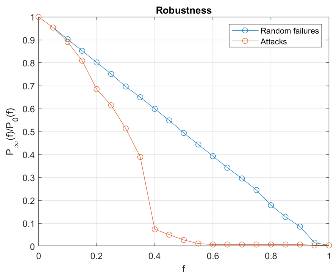
Figure 9: Clustering coefficients on the bipartite matrix.

the deletion of ingredients that are not hubs does not affect too much the size of the giant component.

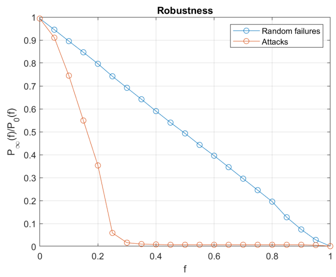
The three networks have **different behaviours as regards the robustness to attacks**.
Particular relevance assumes the *Taiwanese network* where *ingredients are strongly connected with hubs*, and removing one hub has a strong impact on the giant component size. Its curve has consequently *higher slope* than the other two.
In Italy and Japan, the network is still connected until the deletion of the 40% of nodes, while in Taiwan this percentage drops to 20%.



(a) Italy



(b) Japan



(c) Taiwan

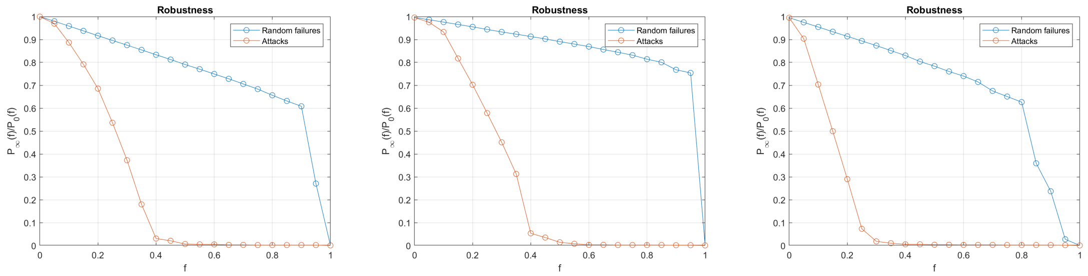
Figure 10: Robustness of projection networks.

	Italy	**Taiwan**	**Japan**
Inhomogeneity ratio κ_f	2705.1593	2576.9876	2744.1992
Breaking point f_c	0.99965	0.99963	0.99965

Table 5: Robustness parameters.

Figure 11 shows that in the **bipartite** networks *the curves relative to attacks assume a similar behaviour as in the projection*, while *the curves relative to random failures are different.*

Their behauviour suggest that the links between recipes and ingredients make the network more connected and robust to random failures.
As expected, the size of the giant component remains very large until the deletion of the 90% of nodes for Italy and Japan, and the 80% for Taiwan, to descrease drastically once ovecame that percentage.

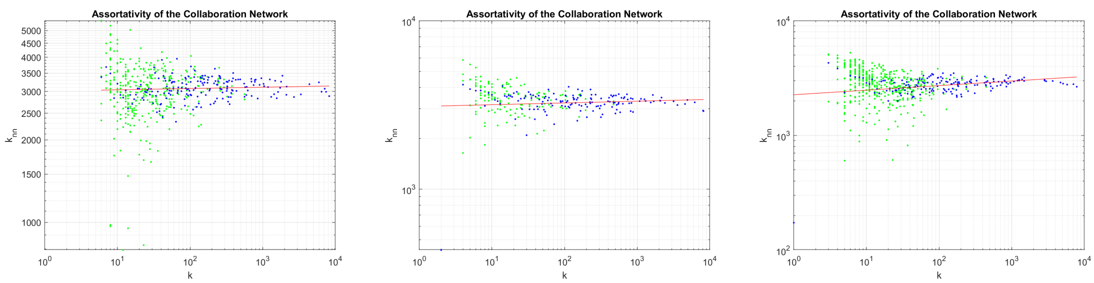


(a) Italy (b) Japan (c) Taiwan

Figure 11: Robustness of bipartite networks.

2.5 Assortativity

The **projection** network is here invesigated.
From figure 12 we can assume that all the networks are *not very assortative*.
However the Taiwanese network assumes a little more assortative behaviour, this explains the results obtained in section 2.4, in particular this is the reason why this network is the least robust to targeted attacks.
Table 6 resumes the most meaningful results. Italian and Japanese networks are **neutral networks**, as their assortativity values are very close to zero.
Taiwan has a little higher assortativity value, but still very low.



(a) Italy (b) Japan (c) Taiwan

Figure 12: Assortativity.

	Italy	**Taiwan**	**Japan**
Natural cutoff	59056.7835	6974.3789	5185.6169
Assortativity value	0.0051636	0.039778	0.010466

Table 6: Assortativity parameters.

3 Second Part

In this section we are going to discuss some *social* features of the Network.

3.1 Ranking

The main idea behind Ranking is to find the *authorities* (nodes with the highest number of incoming links) and the *hubs* (nodes with the highest number of outgoing links). However our networks are **undirected**, so there are no outgoing or incoming edges.
Anyway the analysis were performed by means of *PageRank* and *HITS* algorithms both on the bipartite and projection matrix.

- **PageRank** is based on the equation:

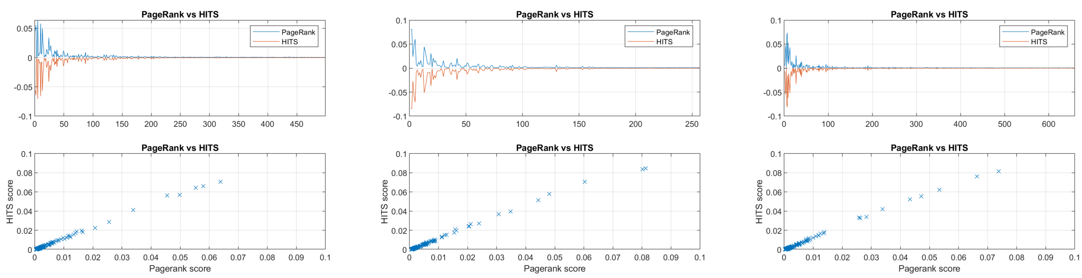
$$\boldsymbol{p_{t+1}} = cM\boldsymbol{p_t} + (1-c)\boldsymbol{q}, \qquad M = A \cdot diag^{-1}(\boldsymbol{d})$$

  where $\boldsymbol{d}$ is the degree vector, $c = 0.85$ is the *dumping factor* and $\boldsymbol{q}$ is the *teleport vector*, set to $\boldsymbol{q} = \frac{\mathbb{1}}{N}$ ($N =$ nodes) on the projection matrix and to $\frac{[\mathbf{1}_k, \mathbf{0}_{N-k}]}{N}$ or $\frac{[\mathbf{0}_k, \mathbf{1}_{N-k}]}{N}$ respectively to emphasize ingredients (first k positions) or recipes (last $N-k$ positions).
  The authorities are given by $\boldsymbol{r} = \boldsymbol{p_\infty}$ and the solution can be found solving a linear system through power iteration. We tried with both.

- **HITS** is based on the equation:

$$\boldsymbol{a_{t+1}} = M\boldsymbol{a_t}, \qquad M = AA^T$$

  where $\boldsymbol{a}$ are the authority scores. HITS was performed only on the projection matrix.

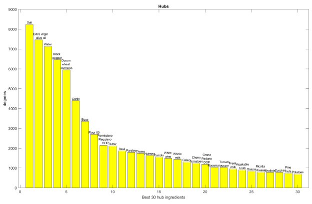


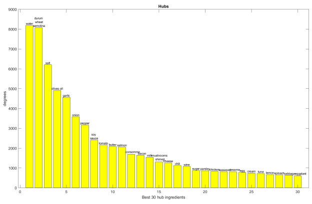
(a) Italy (b) Japan (c) Taiwan

Figure 13: Pagerank and HITS algorithms on the projection matrices. The linear interdependence between them confirms the results obtained.

The result is that *the hubs corresponds to the authorities* but there are *some over-takings* in the lower degree nodes. It is confirmed by the histogram of figure 15 which shows the top 30 ranked ingredients by the two algorithms, compared with the ingredients with the highest degrees.
As expected the hubs/authorities are ingredients owing to the *dough* for each of the three countries, but also *dressing* ingredients used to season many pasta dishes.

In addition we performed the ranking also using **SimRank** algorithm and we built the matrix, whose column correspond to the different $\boldsymbol{p}$ vectors for each node, to use it then in the link prediction.


(a) Italy


(b) Japan

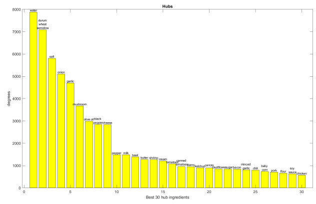
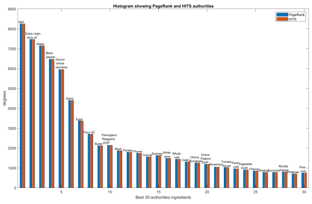
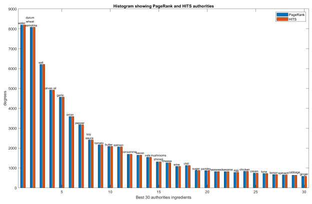

(c) Taiwan

Figure 14: Top 30 hubs.


(a) Italy


(b) Japan

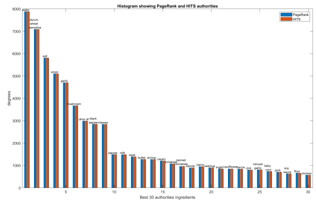

(c) Taiwan

Figure 15: PageRank and HITS best 30 results.

3.2 Communities

The analysis of the communities are performed only on the projection matrix. Many algorithms for this kind of analysis exist. We chose to implement **Spectral Clustering** (SC) and **Page Rank Nibble** (PRN) on the three matrices.
The resulting partitions of the networks are shown in figures 17 and 21, while the highest degree elements of each networks are shown in figure 19.

The number of communities was estimated through the following *quality measures*:

- The **Conductance** $\Phi(\cdot)$ defined as:

$$\Phi(k) = \min_{|S|=k} \phi(S), \qquad \phi(S) = \frac{cut(S, S^c)}{\min(assoc(S), assoc(S^c))}$$

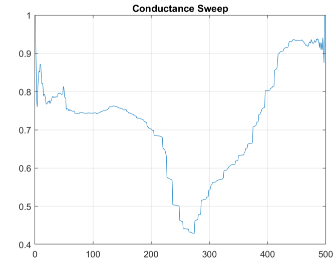
  where S and S^c are the two communities.

- The **Modularity** Q:

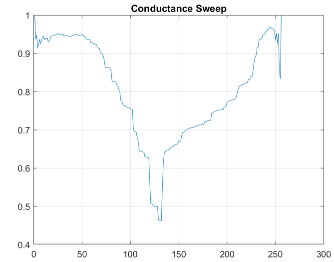
$$Q = \frac{1}{2L} \sum_{i,j} \left(a_{i,j} - \frac{k_i \cdot k_j}{2L} \right) \cdot \eta(c_i = c_j), \qquad \eta = \begin{cases} 1 & if \quad \text{true}, \\ 0 & if \quad \text{false}. \end{cases}$$

  where $a_{i,j}$ are the elements of the projection matrix.

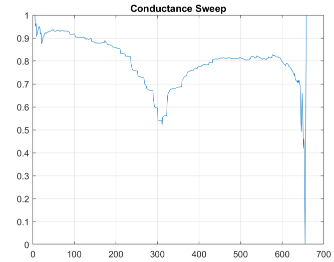
For fist we tried the **K-means** algorithm to find the optimal subdivision looking ant the trend of the modularity and then make it through SC and PRN.
We found that the best subdivision is in *2 communities* for Italy and Japan, in *4 communities* for Taiwan.
The plot of the conductance with a characteristic *V shape* with a single relevant local minimum clearly confirms these results.



(a) Italy

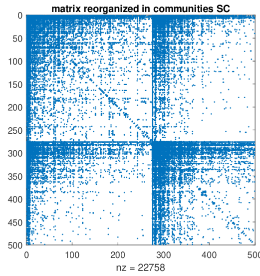


(b) Japan

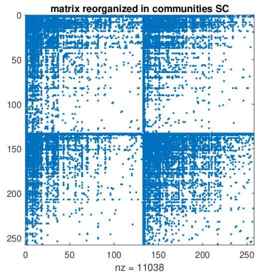


(c) Taiwan

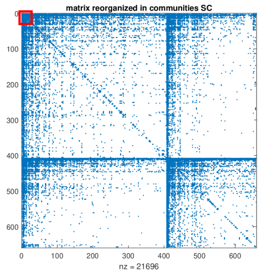
Figure 16: Conductance Sweeps (SC).



(a) Italy

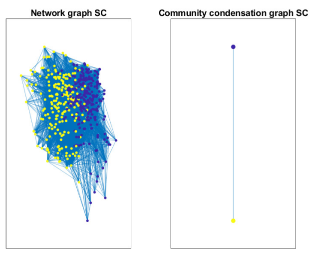


(b) Japan

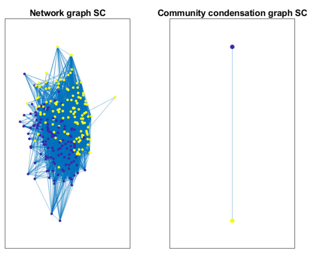


(c) Taiwan

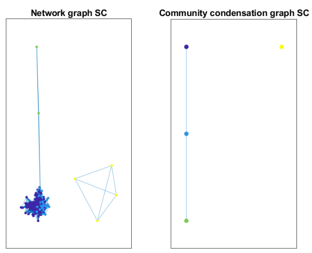
Figure 17: Matrices reorganized in communities (SC).



(a) Italy

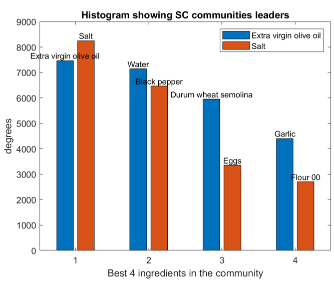


(b) Japan

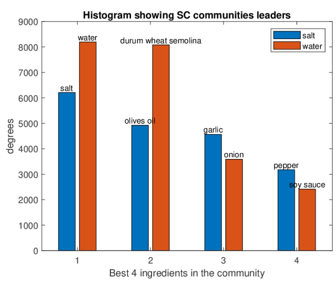


(c) Taiwan

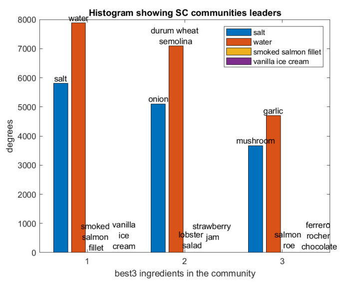
Figure 18: Graph of the networks reorganized in communities on the left and Community condensation graph on the right (SC).



(a) Italy



(b) Japan



(c) Taiwan

Figure 19: Histogram of the highest rank ingredients of each community (SC).

	Italy	**Taiwan**	**Japan**
Modularity Q	0.101	0.048	0.071

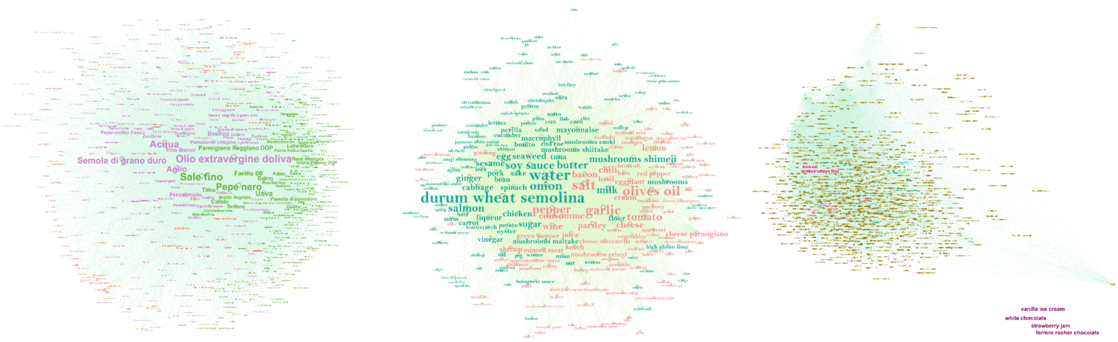
Table 7: Modularity highest values for the three networks. They are all *suboptimal* ($\notin [0.3, 0.7]$).

For **Italy** and **Japan** both Spectral Clustering and Page Rank Nibble subdivide the network into *two communities*. SC makes an equal subdivision of the nodes, while PRN tends to *separate the dough* aside w.r.t. other ingredients.

Taiwan network, on the other side, assumes a different behaviour because its optimal level of subdivision results being of *4 communities* both for PRN and SC.
The first community separated both from SC and PRN is the *disconnected* one, the second subdivision separates the *whiskers* and the final one cuts the giant component. The subdivision is performed by successive bipartitions.

3.2.1 Additional Results

Additionally, we tried to split the network into communities and measuring some performances relying on **Gephi tool**.
The result given exploits a different modularity algorithm[1] and and give us an *additional subdivision* into communities (3 for Italy and Japan, 7 for Taiwan) however recognizing the ones found by SC and PRN.



(a) Italy (b) Japan (c) Taiwan

Figure 20: Subdivision in communities.

Finally a tentative to find overlapping community is done. The algorithm chosen is **K-cliques** where K refers to the minimum size of the cliques. The result is a heap of little communities overlapping each other, but apparently with no specific meaning.

3.3 Link Prediction

These analyses were performed via different algorithms exploiting different features of the network, both on Bipartite and Projection matrices, which establish how two nodes are likely to link according to a similarity matrix $\boldsymbol{S}$ (table 10).

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000

The various algorithms were applied on the *test set* ($T = 90\%$ of the nodes) and measured on the *probe set* ($P = 10\%$ of the nodes). Specifically we took into account the following measures:

- The **Area Under the ROC curve** (AUC) defined as:

$$AUC(P) = \sum_{p \in P, i \in I} \frac{\eta(\boldsymbol{S}(p) > \boldsymbol{S}(i))}{|P| \cdot |I|}, \qquad \eta = \begin{cases} 1 & if \quad \text{true,} \\ 0 & if \quad \text{false.} \end{cases}$$

  where I is the set of *inactive edges* and $\eta(\cdot)$ is a Boolean function.

- The **Precision**, i.e. the percentage of top L links, ranked according to the similarity measure, that belong to the probe set P. We chose $L = 100$.

In order to obtain better results, after applying Link Prediction algorithms we filter results *deleting the ingredients of dough* (i.e. *semolina*, *flour* and *water*). Tables 8 and 9 reports the best matches obtained as the best and most recurrent results given by all the algorithms.

3.3.1 Results

As regards the **Bipartite Matrix**, the aim is to find predictions on *a possible ingredient to add to the recipes.* Some of the best pairings are reported in table 8. From table 10 we can conclude that the *Local community degree* set of algorithms give better results in term of performances.

As regards the **Projection Matrix**, the aim is to find predictions on *possible combinations of ingredients.* Table 10 reports the obtained results.
We can notice that:

- For **Italy** the new pairings are very uncommon for our culture (e.g. *Whole milk-Onions*) but some of them could be very tasty. The algorithms give similar results made exception for RA and AA, which likes *Pig cheecks.*

- For **Taiwan** we can state that some pairings are very uncommon and with *strong flavors.* AA privileges the *carrots.*

- For **Japan** The pairings are very reasonable and appetizing. The different algorithms give similar results, made exception for RA and AA. In particular AA best results always include *salt.*

Also in term of quality measures the algorithms are very similar. AA and RA achieves little worst performances.

3.3.2 Robustness of new recipes

In order to test the new built recipes, a measure of *robustness* was performed on the matrices with the new links.
We take into consideration the following matrices:

- A matrix built simply *adding the best new link* for each recipe.

- A matrix built adding the best new link and *removing the most similar ingredient* to the one added, according to a *similarity measure* ($\boldsymbol{S}$).

New Ingredient	Recipe
Black pepper	Durum wheat semolina, Water, Ricotta salata, Eggplant, Garlic, Vine-ripened tomatoes, Basil, Salt, Extra virgin olive oil
Vegetable broth	Semolina durum whole wheat, Water, Fresh onion, Mushrooms, Bacon, Cannellini beans, Rosemary, Extra virgin olive oil, Black pepper, Salt
apple	onion, anchovies, water, olive oil
Brandy	Chicken breast, Noodles, Potatoes, Snow peas, Carrots, Celery, Mushrooms, Leeks, Water, Fresh ginger, Parsley, Extra virgin olive oil, Black pepper, Salt
Almonds	streaky pork, durum wheat semolina, water, minced garlic, plum, cauliflower, mushroom, soft-boiled eggs, rice wine, salt, flour

(a) Italy

New Ingredient	Recipe
mushroom	onion, meat, red wine, concentrated tomato paste, chicken broth, bay leaves, sugar, salt, durum wheat semolina, water, cheese, fresh thyme, black pepper
chia	streaky pork, durum wheat semolina, water, minced garlic, plum, cauliflower, mushroom, soft-boiled eggs, rice wine, salt, flour
cheese	durum wheat semolina, water, bacon, asparagus, shrimp, garlic, black pepper, rose salt, paprika, parsley leaf, cheese
basil leaves	durum wheat semolina, water, onion, cream, chicken breast, squid
avocado	durum wheat semolina, water, bacon, large tomatoes, green pepper, mushroom, cheese, ketchup, salt, black pepper

(b) Taiwan

New Ingredient	Recipe
consomme	durum wheat semolina, water, salmon, olives oil
tomato	onion, bacon, garlic, olives oil, cream, salt, cheese, durum wheat semolina, water, juice, nut
soy sauce	chicken, salt, durum wheat semolina, water, avocado, clams, mayonnaise, onion, cod roe
onion	durum wheat semolina, water, saury, salt
pepper	durum wheat semolina, water, salmon, olives oil

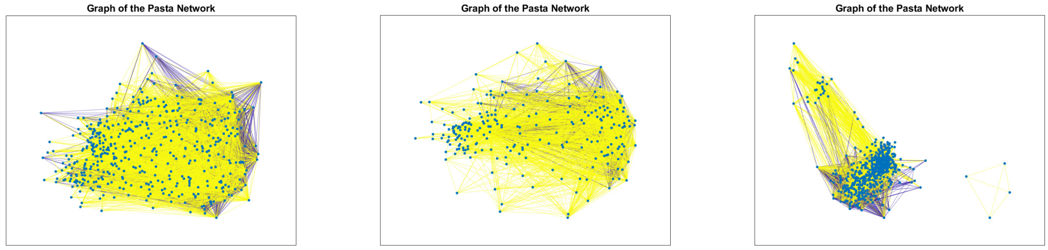
(c) Japan

Table 8: Some of the most interesting additions, obtained from different algorithms.

The analysis is repeated taking into account different similarity measures (e.g. *Common Neighbours*, *Katz* etc.). The result is shown in figure 23.
The recipes with the new links seems to be *more robusts to attacks* w.r.t. the original recipes, especially the ones with *the replaced ingredients*. However a random removal attack destroys little quicker the new recipes matrices.

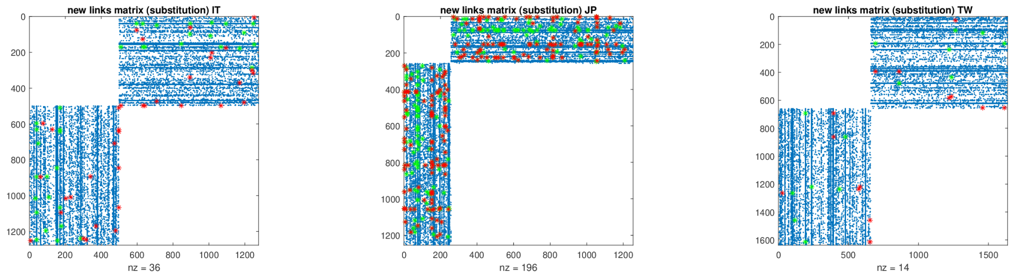
- In **Japan** the most common substitution is with *nut* (i.e. substituting *mushrooms* with: *nut*) and *tomato sauce* can be substituited by *potesara* (A kind of *potato salad*).
- In **Taiwan** the preferred substitutions are with *aivar*, an Eastern sauce.
- In **Italy** many substitutins are on the spices, i.e *black pepper* or *basil*.

We can therefore conclude that *some substitutions can be made* without damaging the integrity of the network.



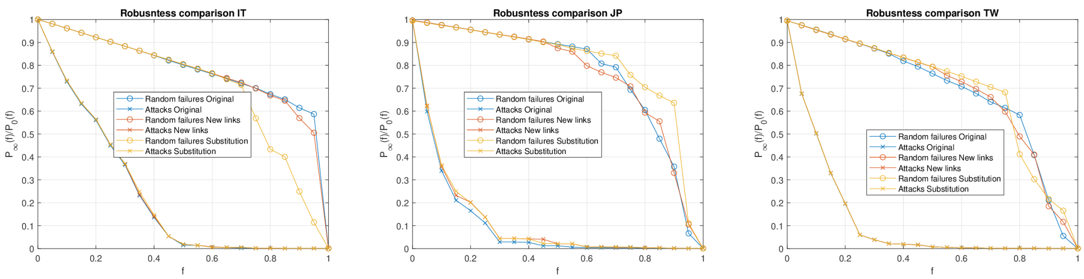
(a) Italy (b) Japan (c) Taiwan

Figure 21: Graph of the networks with the new added links (purple).



(a) Italy (b) Japan (c) Taiwan

Figure 22: The three bipartite matrices with the new added links (red) and the most similar to them, replaced (green). Zero values are avoided and the result is that in Japan the substitutions are even more.



(a) Italy (b) Japan (c) Taiwan

Figure 23: Robustness of new links. (RW similarity based).

pairings		**CN**	**AA**	**RA**	**KA**	**LP**	**RW**
Nutmeg	Fresh chilli	x			x	x	
Liquid fresh cream	Carrots	x			x	x	
Tomato sauce	Pine nuts	x			x	x	
Butter	Mussels	x			x	x	
Salt	Nduja						x
Pig cheek	Pumpkin		x				
Pig cheek	Ricotta cheese	x					
Sausage	Pecorino			x			
Whole milk	Beans			x			
Whole milk	Onions golden		x		x	x	

(a) Italy

pairings		**CN**	**AA**	**RA**	**KA**	**LP**	**RW**
fresh cream	chili	x		x	x	x	
black pepper	potato	x					
spices	bacon	x			x	x	
carrots	nuts		x				
canned tomatoes	pesto	x			x	x	
carrots	pesto		x				
salt	pig cheek						x
lemon juice	chicken broth		x				
rosemary	chicken broth			x			
fresh cream	sugar	x		x	x	x	

(b) Taiwan

pairings		**CN**	**AA**	**RA**	**KA**	**LP**	**RW**
cheese	sesame	x			x	x	
macrophyll	bean			x			
salt	sweet sauce		x				x
cabbage	lemon			x			
lemon	mushrooms maitake			x			
chicken	vegetables			x			
cabbage	cheese parmigiano			x			
consomme	perilla	x			x	x	
egg	lemon	x		x	x	x	
bacon	vinegar	x			x	x	

(c) Japan

Table 9: Best 10 coupling results, obtained taking best results from the different algorithms.

Algorithm	**AUC**	**Precision**
Common Neighbours CN	0.754661	0.000000
Adamic Adar AA	0.755590	0.000000
Resource Allocation RA	0.758653	0.000000
Common Neighbours CAR	0.999967	0.000000
Adamic Adar CAA	0.999835	0.000000
Resource Allocation CRA	0.999835	0.000000

(a) Italy Bipartite

Algorithm	**AUC**	**Precision**
Common Neighbours CN	0.925552	0.100000
Adamic Adar AA	0.831491	0.080000
Resource Allocation RA	0.787285	0.120000
Katz ($\beta = 0.85$)	0.915287	0.060000
Local path ($\beta = 0.85$)	0.915266	0.050000
Random Walk RW	0.974741	0.010000

(b) Italy Projection

Algorithm	**AUC**	**Precision**
Common Neighbours CN	0.758894	0.000000
Adamic Adar AA	0.760047	0.000000
Resource Allocation RA	0.764110	0.010000
Common Neighbours CAR	0.999850	0.000000
Adamic Adar CAA	0.999246	0.000000
Resource Allocation CRA	0.999246	0.010000

(c) Taiwan Bipartite

Algorithm	**AUC**	**Precision**
Common Neighbours CN	0.949459	0.080000
Adamic Adar AA	0.883111	0.050000
Resource Allocation RA	0.832137	0.080000
Katz ($\beta = 0.85$)	0.939287	0.080000
Local path ($\beta = 0.85$)	0.939074	0.070000
Random Walk RW	0.985023	0.010000

(d) Taiwan Projection

Algorithm	**AUC**	**Precision**
Common Neighbours CN	0.880564	0.080000
Adamic Adar AA	0.880365	0.090000
Resource Allocation RA	0.881276	0.020000
Common Neighbours CAR	0.998854	0.000000
Adamic Adar CAA	0.998805	0.090000
Resource Allocation CRA	0.998805	0.020000

(e) Japan Bipartite

Algorithm	**AUC**	**Precision**
Common Neighbours CN	0.941640	0.008000
Adamic Adar AA	0.755516	0.120000
Resource Allocation RA	0.794601	0.070000
Katz ($\beta = 0.85$)	0.938240	0.080000
Local path ($\beta = 0.85$)	0.937842	0.080000
Random Walk RW	0.960592	0.000000

(f) Japan Projection

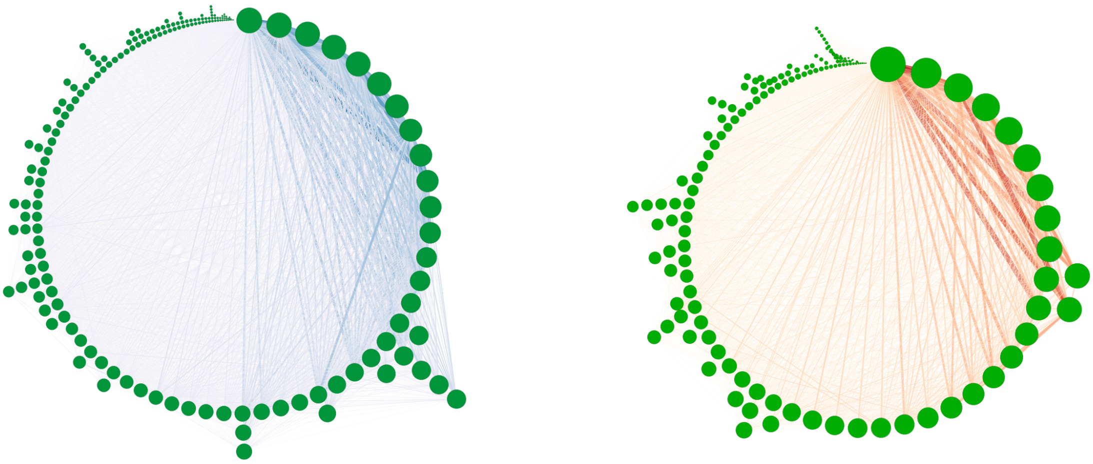
Table 10: On the left the measures for the bipartite matrix, on the right for the projection. The first one takes into account only techniques based on *common neighbours*, while the second both techniques based on *common neighbours*, *path* and *random walk*. The measures have been repeated several times and the results refer to mean values.

4 Noodles

In this section we are going to briefly compare the results obtained analyzing the Taiwanese and Japanese noodles networks with the pasta ones.
We will cover some of the previous analyses and make the point about any similarities/dissimilarities.

4.1 Diameter and distance distribution

The graph of the two networks are here reported. We can notice that, compared to the pasta networks of the corresponding country, *the noodles networks present more high degree ingredients and less weighted edges.*



(a) Taiwan (b) Japan

Figure 24: Noodles graphs.

As far as the diameter is concerned, the result (tab. 11) shows that there is *almost no difference* between the pasta and noodles networks of Japan and Taiwan.

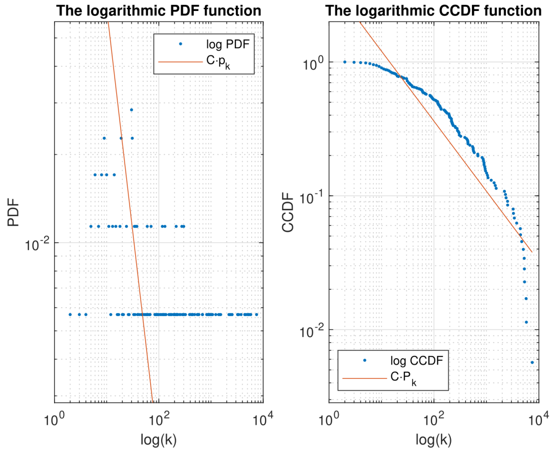
	Taiwan	**Japan**
diameter	5	5
average distance	2.1136	2.1436

Table 11: Diameter and average distance.

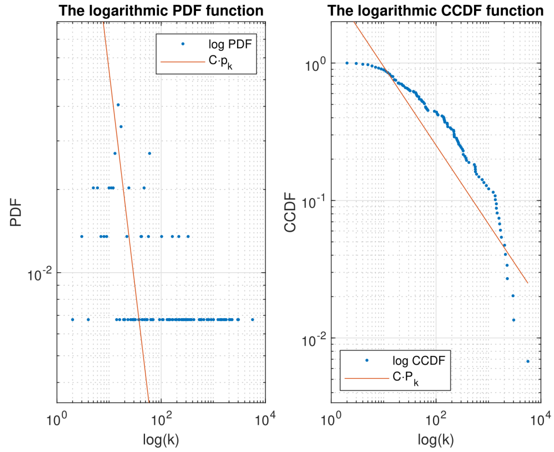
4.2 Network model

Unlike the pasta network, the *Taiwanese noodle network holds the highest average degree* and consequently its power-law exponent γ *is the lowest one*. Both the networks have $\gamma \leq 2$ and do not hold the scale-free behaviour.

The very large presence of high degree nodes in the both networks can be seen in the logarithmic Complementary cumulative density function (CCDF) plot, decreasing more slowly than the CCDF plots of the pasta networks.



(a) Taiwan



(b) Japan

Figure 25: Probability density function (PDF) and Complementary cumulative density function (CCDF).

	Taiwan	**Japan**
Average degree $\langle k \rangle$	645.0341	357.6216
Second Moment $\langle k^2 \rangle$	1930432.0432	163774.423
Third Moment $\langle k^3 \rangle$	45007707020.7499	1188538161.4441
Variance σ^2	1814470.8852	538694.5055
k_{max}	7475	5661
k_{min}	2	2
γ	1.5233	1.5726
γ_{sat}	1.6821	1.7212

Table 12: Other network parameters.

4.3 Robustness

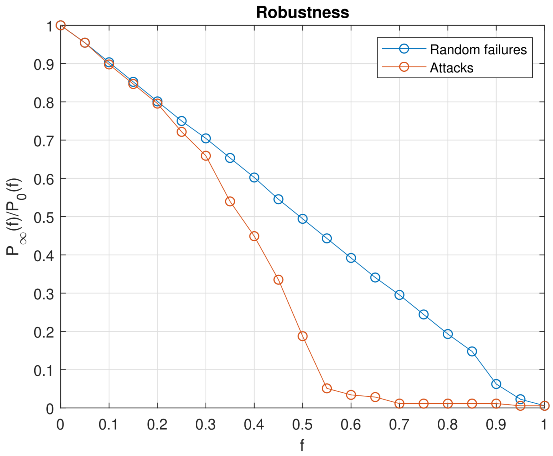
The two noodles networks, compared with the pasta ones, are *more resistant to attacks* (fig. 26).
The networks don't collapse right away as it happened in the pasta ones and this is due to an **higher break-up threshold**.

	Taiwan	**Japan**
Inhomogeneity ratio κ_f	2672.3853	1431.0591
Breaking point f_c	0.99964	0.99934

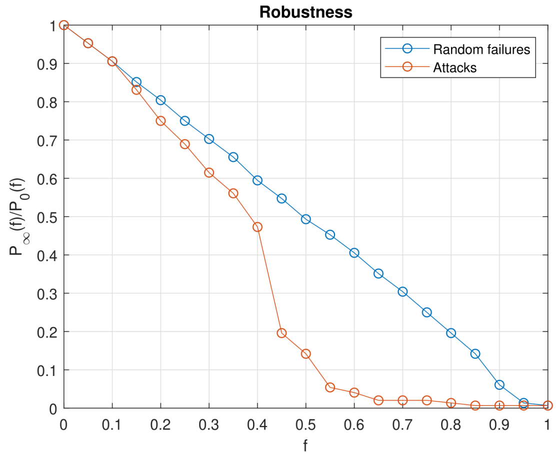
Table 13: Robustness parameters.

4.4 Assortativity

We can here notice a *substantial difference between the two networks.*
The **Taiwanese** network tends to be *slightly more assortative* than the Japanese one, because the most recurrent ingredients (i.e. hubs) tend to be more closely matched, while in Japan we almost see a neutral network with no evidence on a trend on how nodes are wired.
Comparing pasta and noodles we can see a similar trend for both countries, with Taiwan being more assortative and *Japan presenting an almost neutral network behaviour.* For a more detail comparison see 27 and 14

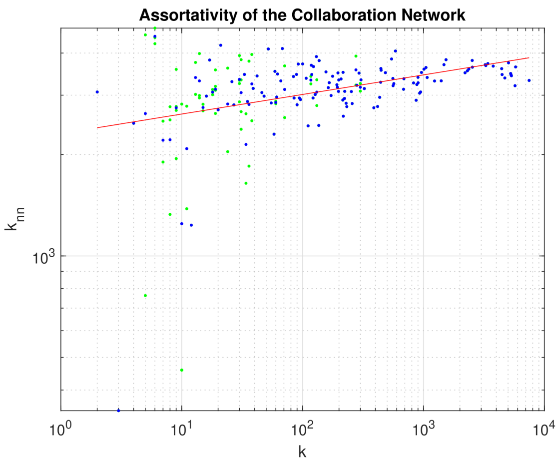


(a) Taiwan

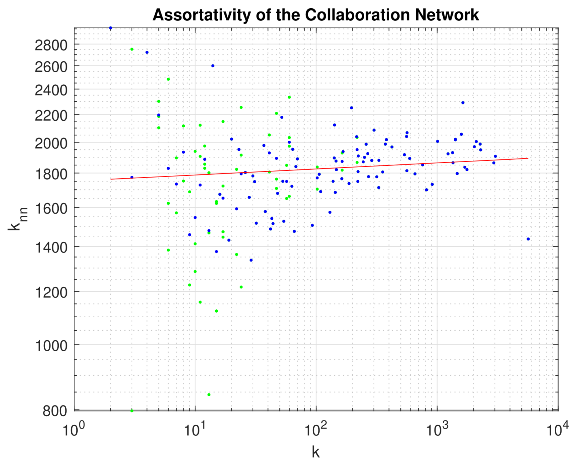


(b) Japan

Figure 26: Robustness.



(a) Taiwan



(b) Japan

Figure 27: Assortativity.

	Taiwan	**Japan**
Natural cutoff	39103.134	12341.2277
Assortativity value	0.058106	0.0089898

Table 14: Assortativity parameters.

4.5 Link Prediction

Analyzing the projection matrix of both Taiwan and Japan the aim is to find out what can be new pairings among the ingredients.

- For **Taiwan** we can state that some combinations are bizarre and resulting in *strong flavors*. Most of the algorithms privilege the use of *shiitake mushrooms* .
- For **Japan** The pairings are very reasonable and appetizing. The different algorithms give similar results, made exception for RA and AA. The most recurring ingredient is *chicken*.

pairings		CN	AA	RA	KA	LP	RW
shiitake	lemon	x			x	x	
shallot sauce	shiitake		x				
cherry	pickle			x			
pork	bacon	x			x	x	
cherry	meat			x			
shiitake	ketchup	x			x	x	
udon	parsley			x			
sacha sauce	buckwheat						x
shallot sauce	pork		x				
cheese	chop			x			

(a) Taiwan

pairings		CN	AA	RA	KA	LP	RW
chicken	mayonnaise	x			x	x	
noodle	condensed milk		x				
udon	miso soup						x
bean	peas			x			
cabbage	broccoli			x			
stock	chop			x			
chicken	tuna	x			x	x	
carrot	lettuce			x			
olive oil	stock			x			
noodle	pineapple		x				

(b) Japan

Table 15: Best 10 coupling results, obtained taking best results from the different algorithms.

5 Conclusions

To conclude we can summarize briefly the results by *answering to the initial questions*:

- Which are the most popular ingredients used for pasta in different cultures?
- Are these ingredients similar or different?
- How similar is the eastern pasta to western pasta vs eastern noodle?

The **most common ingredients for pasta recipes in all the three countries are the ones related to dough and others basic ingredients** (water, durum wheat semolina, salt, olives oil, garlic, onion, pepper). These ingredients are present in all the three countries and are also the ones with *highest degree*. Analyzing the diagrams of figure 28 we can see that *in Italy and Taiwan almost the 75% of the ingredients used for pasta are not present in other countries* (so they are *typical toppings*). In Japan this percentage drops to the 63%.

To resume we can draw the following conclusions:

- **Italian** pasta has more ingredients in common with Taiwanese pasta than with Japanese pasta.
- **Taiwanese** pasta has almost the same amount of ingredients in common with Italian pasta and Japanese pasta.
- **Japanese** pasta has more ingredients in common with Taiwanese pasta than with Italian pasta.

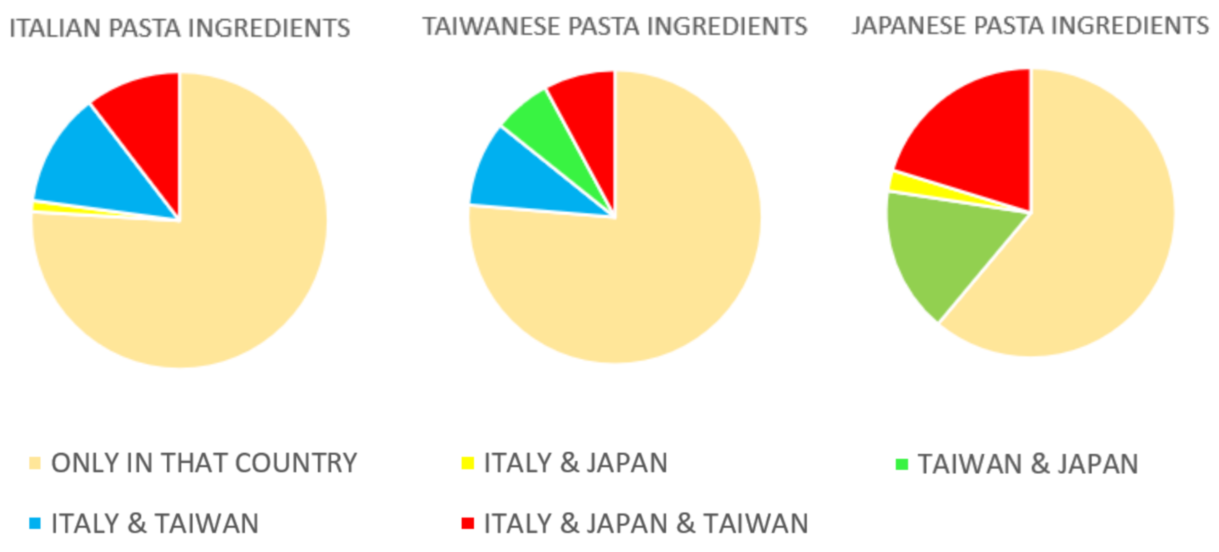


Figure 28: Comparison diagrams.

The *Euler Venn diagram* of figure 29 shows these results.

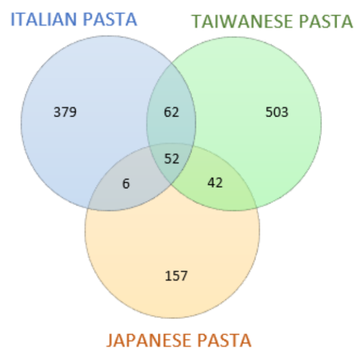


Figure 29: Euler Venn Pasta-Pasta comparison plot.

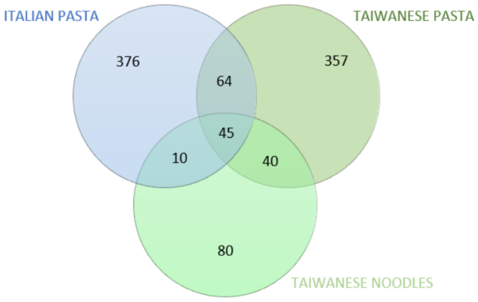
Figure 30 shows a **comparision between western pasta and eastern noodles**.
Figure 30a represents *italian pasta vs taiwanese pasta and taiwanese noodles*. We can notice that:

- **Italian pasta** has more ingredients in common with Taiwanese pasta than with Taiwanes noodles.
- **Taiwanese pasta** has more ingredients in common with Italian pasta than with Taiwanes noodles.
- **Taiwanese noodle** has more ingredients in common with Taiwanese pasta than with Italian pasta.

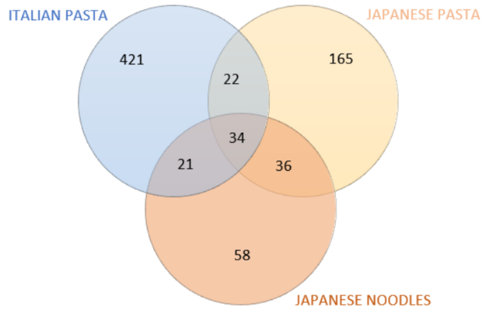
On the other hand figure 30b represents *italian pasta vs japanese pasta and japanese noodles*. We can notice that:

- **Italian pasta** has almost the same amount of ingredients in common with Japanese pasta and Japanese noodles.
- **Japanese pasta** has more ingredients in common with Japanese noodles than with Italian pasta.

- **Japanese noodle** has more ingredients in common with Japanese pasta than with Italian pasta.



(a) Taiwan



(b) Japan

Figure 30: Euler Venn Pasta-(Pasta/Noodle) comparison plot.