Research Report #2 - Generation of soft-pruned PFNet maps

A. Quirin - February 28, 2007

Abstract

In this document, we describe some attributes that can be obtained from full *Visual Science Maps* and pruned *Visual Science Maps*. These attributes can be used as metrics to improve the pruning based on PFNet. We also describe in this document three filters based on these attributes, and their results are discussed.

1 Introduction

For some specific *Visual Science Maps* of the first level (large domain maps), the PFNet pruning algorithm lost some interesting links (edges) because the weights of some others are slightly above them. These links, called *weird links*, does not occur in all the maps.

This is a short summary of the known properties of these links:

- Usually, these links have a low weight (<12, <8 or <4).
- This problem occurs only on countries with a low scientific production, a third-world or an undeveloped country (the ones we have are: Mexico, Argentina, Chile, Cuba, Peru and Venezuela).
- More bad links are observed on short time period (below 1 year), but not on developed countries (for instance, UK).

If we want to discard some links and include some others, it could be usefull to base our filtering on some attributes. Attributes are available for links, nodes and *s. These are described in the next section.

2 Available attributes

Terms: in our case, we work on the first level, so the *domain* of each map corresponds to a whole country. Each node represents a scientific *category*.

2.1 Link-based attributes

These are the attributes available for a link L_{ij} :

The weight (W_{ij}). In our case, it is the sum between the raw cocitation value and the normalized cocitation value. A link with a higher value means that the connection between the two categories has a sense.

The fields of the connected nodes (F_i, F_j). Each category is assigned to a field by the ISI classification. These are the labels of the fields of the two connected nodes.

The categories of the connected nodes (C_i, C_j). These are the labels of the two connected nodes.

The SFL flag ($\mathbf{SFL}_{ij}$). The *Same Field Link* (SFL) flag is set to the value *true* if the connected nodes belong to the same field, and *false* if not. We call the opposite *Different Field Link* flag (DFL). SFL links could be viewed as more meaningful than DFL links. It appears that in the Argentina map pruned by PFNet, 50% of the links are SFL links.

The centrality index ($\mathbf{CEN}_{ij}$) Used to describe the location of the link within the whole map. This is the number of links between L_{ij} and the center of the map, on the shortest path. As the center of the map corresponds to the most important category, the proximity with the center could be viewed as a measure of *importance* of each link.

The border index ($\mathbf{BOR}_{ij}$) Used to describe the location of the link within the whole map. This is the number of links between L_{ij} and the nearest node located in the border of the map. As the pruning remove rather links in the border of a map, this attribute could be seen as a measure of *weakness* of a link.

It could be interesting to see if there is a relationship between the weights of the links and the SFL flag. For instance, we can expect that links with large weights are SFL and links with small weights are DFL. This study can be done with a full and a pruned map. On the Fig. 1, we have printed the rate of the SFL links as a function of their ranks when these links are sorted by their weight values, for the pruned Argentina map. We observe that this function increases linearly. This means that there is absolutly no relationship between the weight of a link and its SFL flag, at least on a PFNet map. So, it is possible that the weight and the SPL flag are two complementary attributes.

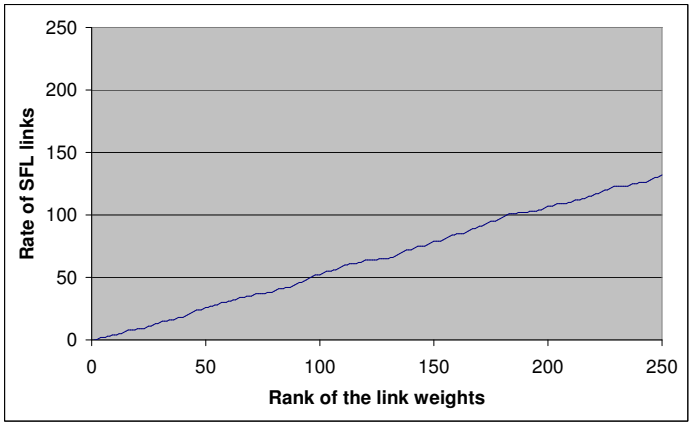

Figure 1: Rate of SFL links VS their ranks for the pruned Argentina map.

We have also studied the relationships between the location of the links in a map and their SFL flags. On the Fig. 2, we have printed a PFNet-pruned Argentina map where the values of the links were set to 1000 if the link is a SFL link, and to 0 if the link is a DFL link. The values of the links are printed in gray, the indexes of the nodes are printed in black. It seems that there is no more relationship between the location of the links and the SFL flag, because the SFL links can appear at any location on the map.

2.2 Node-based attributes

These are the attributes available for a node N_i :

The field (F_i).

The category (C_i).

The number of inner links (L_i). Frequently, the node with the highest number of inner links appears in the center of maps. This could be used to characterize the *importance* of each node.

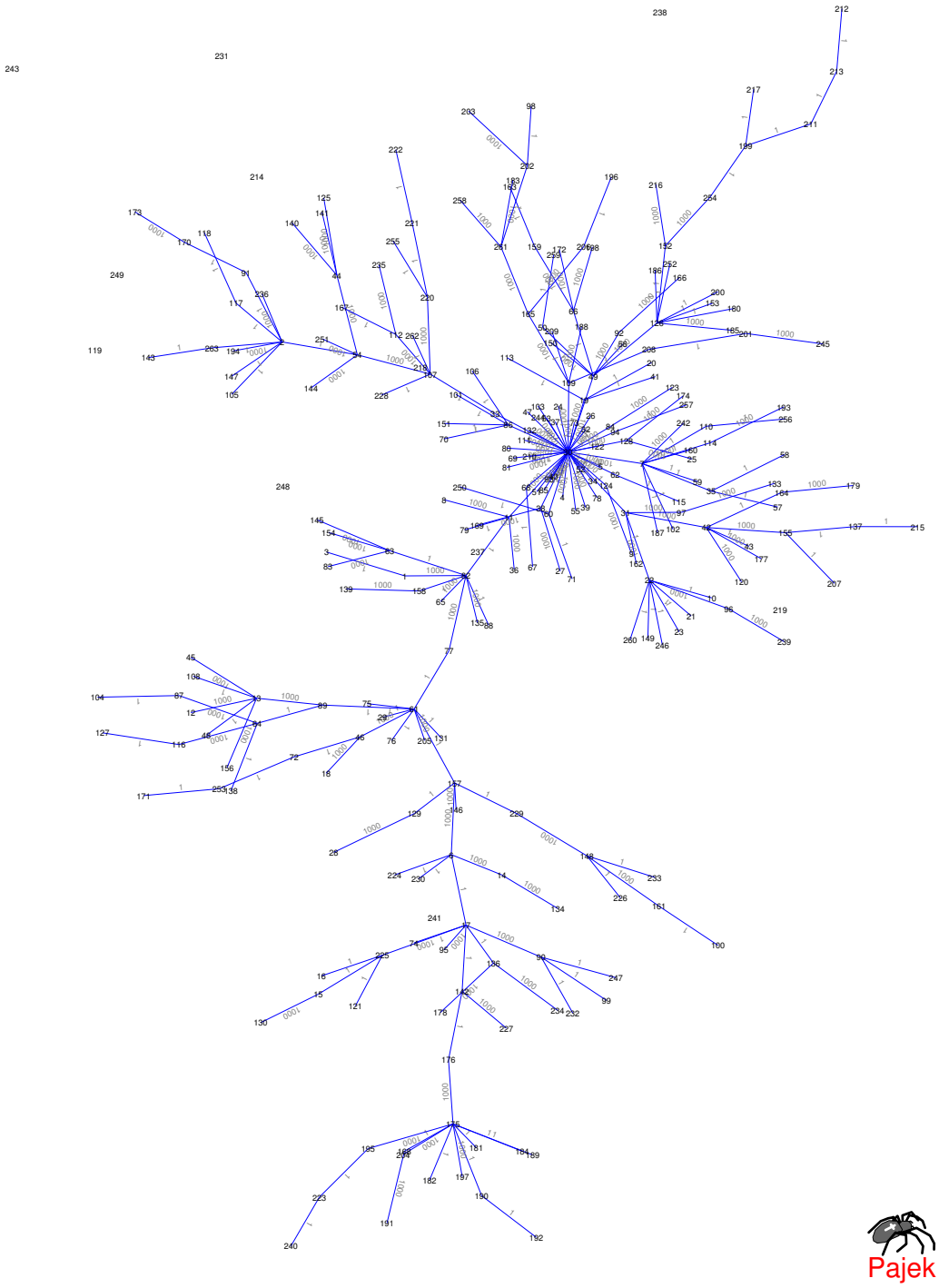

Figure 2: Location of SFL links (W=1000), for the pruned Argentina map.

The number of the inner SFL and DFL links ($L_{\mathbf{SFL}_i}$, $L_{\mathbf{DFL}_i}$). This parameter could be useful to characterize the nodes (is a node connecting two different fields or not?). Note that L_{SFL_i} +

$L_{\mathrm{DFL}_i} = L_i$.

The minimum, maximum and average inner link weights (m_i, M_i, μ_i).

The minimum, maximum and average SFL link weights ($m_{\mathbf{SFL}_i}$, $M_{\mathbf{SFL}_i}$, $\mu_{\mathbf{SFL}_i}$).

The minimum, maximum and average DFL link weights ($m_{\mathbf{DFL}_i}$, $M_{\mathbf{DFL}_i}$, $\mu_{\mathbf{DFL}_i}$). The comparison of these attributes to the previous ones (m_{SFL_i}, M_{SFL_i}, μ_{SFL_i}) could be used to have a more fuzzy way to characterize the nodes connecting two different fields.

The category of the relationship between the minimum and the maximum weights of SFL and DFL links ($C_{\mathbf{WL}_i}$). This attribute is processed using the attributes m_{SFL_i}, M_{SFL_i}, m_{DFL_i} and M_{DFL_i}. For a given node, we can describe the relationship between the minimum and the maximum weights for SFL and DFL links. In this case, we obtain six categories described in the Table 1 and in the Fig. 3. For special cases, we can add additional rules. If there is no inner SFL links ($L_{\mathrm{SFL}_i} = 0$), we define $m_{\mathrm{SFL}_i} = M_{\mathrm{SFL}_i} = 0$. If there is no inner DFL links ($L_{\mathrm{DFL}_i} = 0$), we define $m_{\mathrm{DFL}_i} = M_{\mathrm{DFL}_i} = 0$. Thus, we can apply the constraints defined in Table 1 for all the cases.

Condition	Value of $C_{\mathbf{WL}_i}$
$M_{\mathrm{DFL}_i} < m_{\mathrm{SFL}_i}$	1
$\begin{cases} m_{\mathrm{DFL}_i} < m_{\mathrm{SFL}_i} \\ m_{\mathrm{SFL}_i} \leq M_{\mathrm{DFL}_i} \\ M_{\mathrm{DFL}_i} < M_{\mathrm{SFL}_i} \end{cases}$	2
$\begin{cases} m_{\mathrm{DFL}_i} < m_{\mathrm{SFL}_i} \\ M_{\mathrm{SFL}_i} \leq M_{\mathrm{DFL}_i} \end{cases}$	3
$\begin{cases} m_{\mathrm{SFL}_i} \leq m_{\mathrm{DFL}_i} \\ M_{\mathrm{DFL}_i} < M_{\mathrm{SFL}_i} \end{cases}$	4
$\begin{cases} m_{\mathrm{SFL}_i} \leq m_{\mathrm{DFL}_i} \\ m_{\mathrm{DFL}_i} \leq M_{\mathrm{SFL}_i} \\ M_{\mathrm{SFL}_i} \leq M_{\mathrm{DFL}_i} \end{cases}$	5
$M_{\mathrm{SFL}_i} < m_{\mathrm{DFL}_i}$	6

Table 1: The categories of links, based on the minimum and maximum weights of the SFL and the DFL links.

A special case of the category *5* occurs when the node is isolated, that is $m_{\mathrm{SFL}_i} = M_{\mathrm{SFL}_i} = 0$ and $m_{\mathrm{DFL}_i} = M_{\mathrm{DFL}_i} = 0$. It seems that it is better to use a special category to describe this particular kind of nodes, and we place them in an additional category, the category *7*.

The relationships between DFL/SFL links and their weights seem to play a role during the PFNet pruning. The Table 2 shows the evolution in the distribution of the previously described categories from a full map to a pruned map, for some countries. When we compare the distribution of the nodes within the seven categories, it appears that a PFNet-pruned map put almost all the nodes in categories *1* and *6*. This demonstrates that the PFNet-pruning process tries to avoid overlapping weights for DFL and SFL links connected to the same node. This attribute can be used to describe each node in the same way that the SFL flag, and can also be used to describe more precisely the pruning process of any filter. Note also that there is no node in the categorie *7* for large domain maps.

The centrality index ($\mathbf{CEN}_i$) Used to describe the location of the node within the whole map. This is the number of links between N_i and the center of the map, on the shortest path. This attribute could be viewed as a measure of *importance* of each node.

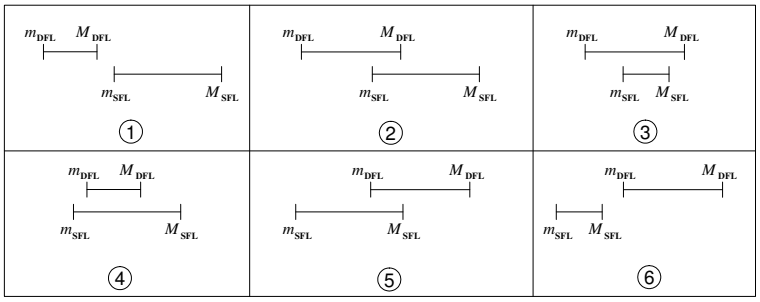

Figure 3: The representations of the C_{WL_i} categories.

Categories:	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Full Argentina map	3	**134**	**105**	1	2	6	12
Pruned Argentina map	**129**	5	9	4	3	**101**	12
Full Chile map	1	**132**	**102**	0	1	4	2
Pruned Chile map	**127**	2	7	4	2	**98**	2
Full Cuba map	5	**111**	**73**	1	3	17	9
Pruned Cuba map	**103**	7	6	7	0	**87**	9
Full Europa map	1	**128**	**88**	0	1	0	0
Pruned Europa map	**118**	8	3	3	3	**83**	0
Full USA map	1	**120**	**96**	0	1	0	0
Pruned USA map	**113**	5	6	3	3	**88**	0

Table 2: Distribution of nodes within the different categories C_{WL_i}, for some maps.

The border index ($\mathbf{BOR}_i$) Used to describe the location of the node within the whole map. This is the number of links between N_i and the nearest node located in the border of the map. This attribute could be seen as a measure of *weakness* of a link.

2.3 Global attributes

These are the attributes available for a map G. Global attributes are important to produce normalized local attributes. They can also be used to build more complexe measures using weighted local-attributes (node-based attributes or link-based attributes). Small and large maps will not have the same average link weights. So, for instance, a thereshold adapted to each particular map could be built using these global attributes.

The number of nodes, links, SFL and DFL links (N_G, L_G, $L_{\mathbf{SFL}_G}$, $L_{\mathbf{DFL}_G}$).

The minimum, maximum and average link weights (m_G, M_G, μ_G).

The minimum, maximum and average SFL link weights ($m_{\mathbf{SFL}_G}$, $M_{\mathbf{SFL}_G}$, $\mu_{\mathbf{SFL}_G}$).

The minimum, maximum and average DFL link weights ($m_{\mathbf{DFL}_G}$, $M_{\mathbf{DFL}_G}$, $\mu_{\mathbf{DFL}_G}$).

3 A filter $\mathcal{F}_1$ based on the weights (W_{ij}) and the SFL flag ($\mathbf{SFL}_{ij}$)

The filter shown in Fig. 4 generates a new pruned network by pre-processing the original network before the PFNet-pruning. It uses a threshold T to eliminate the low-valued links (with a weight below

T) that are also DFL links. In a formal way, to obtain the network NET(4), we delete all links L_{ij} so that :

$$\mathrm{SFL}_{ij} = \mathit{false} \wedge W_{ij} < T$$

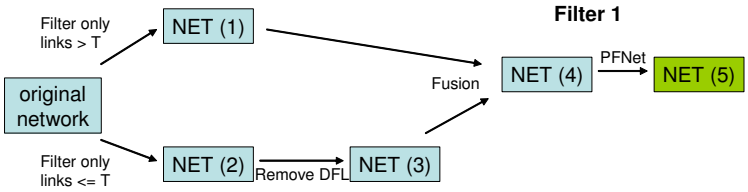

Figure 4: Description of the filter 1.

The results obtained are interesting, in the sense that meaningful links appear and meaningless links are deleted, when compared to a non-filtered PFNet-pruned map. For instance, for the Argentina map,

- Added links (compared to the PFNet map):

```
(Medicine:MEDICINE, LEGAL) <=> (Medicine:MEDICAL ETHICS), weight: 7.165869
(Social Sciences:ANTHROPOLOGY) <=> (Social Sciences:CRIMINOLOGY & PENOLOGY), weight: 1.015694
(Psychology:EDUCATION & EDUCATIONAL RESEARCH) <=> (Psychology:PSYCHOLOGY, EDUCATIONAL), weight: 5.0
(Humanities:PHILOSOPHY) <=> (Humanities:CLASSICS), weight: 1.025079
```

- Removed links:

```
(Geosciences:GEOSCIENCES, INTERDISCIPLINARY) <=> (Humanities:FILM, RADIO, TELEVISION), weight: -5.9
(Medicine:MEDICINE, GENERAL & INTERNAL) <=> (Psychology:EDUCATION & EDUCATIONAL RESEARCH), weight:
(Medicine:MEDICINE, GENERAL & INTERNAL) <=> (Humanities:LITERATURE, BRITISH ISLES), weight: -4.9868
(Agriculture:AGRICULTURAL ECONOMICS & POLICY) <=> (Humanities:CLASSICS), weight: -5.982329
(Social Sciences:ANTHROPOLOGY) <=> (Humanities:FOLKLORE), weight: -6.964907
(Psychology:PSYCHOLOGY, EXPERIMENTAL) <=> (Humanities:MUSIC), weight: -4.970054
(Economics:LAW) <=> (Social Sciences:CRIMINOLOGY & PENOLOGY), weight: -5.905191
(Medicine:HEALTH POLICY & SERVICES) <=> (Social Sciences:*SOCIAL SCIENCES), weight: -5.833333
(Humanities:PHILOSOPHY) <=> (Social Sciences:ETHNIC STUDIES), weight: -6.943923
(Medicine:MEDICAL ETHICS) <=> (Humanities:ETHICS), weight: -0.710586
```

4 A filter $\mathcal{F}_2$ based on a scored DB

In all the networks that were processed, many bad deletions were common to many networks. For instance, the link Economics:LAW <=> Social Science:CRIMINOLOGY was deleted because Economics and Social Science are different fields. Of course, law and criminology are highly related, and this relation should be kept. In fact, this issue comes frequently with the same links, so the amount of *bad deleted links* is small. In this case, the labellisation of each link by hand can overcome this issue.

We have decided to use a score to label such links. An expert is expected to weight some links A-B with three possible values :

- -1 means that the field of A is the same as the field of B, but the two corresponding nodes should be considered as not related (for instance, Humanities:FILM <=> Humanities:RELIGION),
- 0 means that we have to follow the classical ISI database and,
- +1 means that the field of A is different than the field of B, but the two corresponding nodes should be considered as related (for instance, Economics:LAW <=> Social Science:CRIMINOLOGY)

Bad-removed links are scored by the value +1, to approve them and bad-added links are scored by the value -1, to remove them. All these scores are saved in a scored database (DB). The process of $\mathcal{F}_2$ is described in Fig. 5.

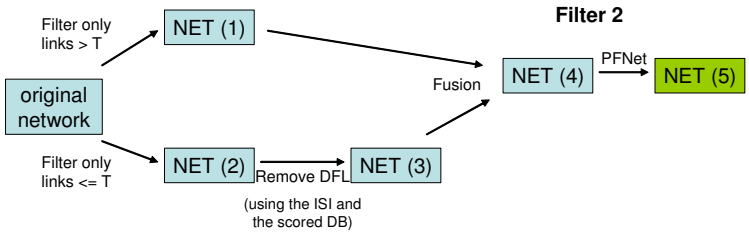

Figure 5: Description of the filter 2.

Using this DB in conjunction with the ISI database to detect SFL/DFL links, and a low threshold (T=4), we obtain this result for the Argentina map:

- Added links:

```
(Humanities:PHILOSOPHY) <=> (Humanities:CLASSICS), weight: 1.025079
```

- Removed links:

```
(Geosciences:GEOSCIENCES, INTERDISCIPLINARY) <=> (Humanities:FILM, RADIO, TELEVISION), weight: -1.9
(Medicine:MEDICINE, GENERAL & INTERNAL) <=> (Humanities:LITERATURE, BRITISH ISLES), weight: -0.9868
(Agriculture:AGRICULTURAL ECONOMICS & POLICY) <=> (Humanities:CLASSICS), weight: -1.982329
(Social Sciences:ANTHROPOLOGY) <=> (Humanities:FOLKLORE), weight: -2.964907
(Psychology:PSYCHOLOGY, EXPERIMENTAL) <=> (Humanities:MUSIC), weight: -0.970054
(Humanities:PHILOSOPHY) <=> (Social Sciences:ETHNIC STUDIES), weight: -2.943923
```

To see the influence of the threshold on the results, we have also compared the results incrementaly, using a variable threshold. In this study, the map obtained using a threshold T is compared to the map obtained using the threshold T+1. Less different are a map to another, more robust is the process. The Table 3 presents the results of the incremental comparison.

According to Table 3, it seems that correct links are added, and bad links are deleted, until a given threshold. After $T = 7$, some links are deleted but should not (for instance, (PSYCHOLOGY) <=> (SOCIAL WORK)). This could suggest the right threshold for this particular case. It seems also that the filter is robust to the threshold parameter, because not so many links are added or suppressed when this threshold is adjusted.

But this filter suffers of some problems:

T	Added links	Deleted links
1		(ANTHROPOLOGY) => (FOLKLORE) (PHILOSOPHY) => (ETHNIC STUDIES)
2	(PHILOSOPHY) => (CLASSICS)	(GEOSCIENCES, INTERDISCIPLINARY) => (FILM, RADIO, TELEVISION) (AGRICULTURAL ECONOMICS & POLICY) => (CLASSICS)
3		(MEDICINE, GENERAL & INTERNAL) => (LITERATURE, BRITISH ISLES) (PSYCHOLOGY, EXPERIMENTAL) => (MUSIC)
4		
5	(LITERATURE) => (RELIGION)	(HISTORY) => (RELIGION)
6		
7	(EDUCATION & EDUCATIONAL RESEARCH) => (PSYCHOLOGY, EDUCATIONAL)	(MEDICINE, GENERAL & INTERNAL) => (EDUCATION & EDUCATIONAL RESEARCH)
8	(SOCIAL WORK) => (FAMILY STUDIES)	(PSYCHOLOGY) => (SOCIAL WORK)

Table 3: Incremental comparison of the maps obtained using $\mathcal{F}_2(T)$ and $\mathcal{F}_2(T+1)$, for the Argentina map.

- The threshold parameter is still here, and some expert knowledge should be used to adjust it. The correct value, even using an incremental method, is difficult to set up because there is no clear separation between a threshold giving only good additions/deletions and a threshold allowing bad links to appear. In fact, good additions/deletions could also appear above a highest threshold, so it is hard to take a decision.
- The new database is more subjective than the first one, even if this gives a better interaction between the expert and the filter, and let the expert have a better control on it.

5 A filter $\mathcal{F}_3$ based on an adaptative threshold

In this filter, we have used the attributes defined previously for the definition of an adaptative threshold (defined without the need of an user). The idea of this filter is to define the value of the new threshold using the weights of the SFL links. SFL links should corresponds to meaningful links, so a good threshold could be above the values of their weights. This value is computed for each node. So, in this filter, we delete all the DFL links having a value smaller than the minimum weight of the SFL links connecting their nodes. In a formal way, we delete all links L_{ij} so that :

$$\mathrm{SFL}_{ij} = \mathit{false} \wedge W_{ij} < m_{\mathrm{SFL}_i}$$

The process of $\mathcal{F}_3$ is described in Fig. 6. This filter is interesting because it uses the value of the *trusted* links (the SFL links) to prune the *untrusted* links (the DFL links). The minimum value of the SFL links is used as the threshold for each node, and this operation is repeated twice because each link to prune has two nodes. Note that working with raw or normalized values would yield to the same results.

But this adaptative threshold seems to include too much freedom in the filter. Some SFL links have high weights, especially in the center of the maps, so good DFL links are then discarded. This problem can be overcome by using an additional parameter to disable the prune when the weights of the links are large (it could be an other threshold or a percentage). But this will require again some expert knowledge.

We can also use the location in the map as an heuristic to enable or disable this filter, but this was not tested now.

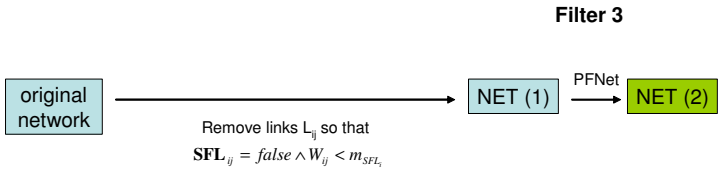

Figure 6: Description of the filter 3.

For the moment, we did not obtained good results with this filter. For instance, this is the results on the Argentina map :

- Added links:

```
(Engineering:ENGINEERING, PETROLEUM) <=> (Humanities:FILM, RADIO, TELEVISION), weight: 1.019024
(Psychology:EDUCATION, SCIENTIFIC DISCIPLINES) <=> (Physics:PHYSICS, MULTIDISCIPLINARY), weight: 28
(Engineering:ENGINEERING, CHEMICAL) <=> (Materials Science:MATERIALS SCIENCE, TEXTILES), weight: 36
(Medicine:MEDICINE, MISCELLANEOUS) <=> (Humanities:CLASSICS), weight: 2.017575
```

- Removed links:

```
(Geosciences:GEOSCIENCES, INTERDISCIPLINARY) <=> (Humanities:FILM, RADIO, TELEVISION), weight: -2.0
(Psychology:EDUCATION, SCIENTIFIC DISCIPLINES) <=> (Chemistry:CHEMISTRY, MULTIDISCIPLINARY), weight
(Agriculture:AGRICULTURAL ECONOMICS & POLICY) <=> (Humanities:CLASSICS), weight: -2.017671
(Chemistry:CHEMISTRY, APPLIED) <=> (Materials Science:MATERIALS SCIENCE, TEXTILES), weight: -50.041
```

6 Conclusion

Because of the difficulty to clearly identify the most important attributes and build a filter especially designed to prune the maps of low scientific production countries, it could be useful to adopt a more systemic approach. In this approach, all the attributes could be analyzed together to detect which ones triggers the prune. A full dataset could be set up in which the maps are decomposed into several attributes. We propose the following table of attributes (see Table 4). A dataset of links, including all the link attributes and the predictive attribute could be processed by some advanced algorithms (neural networks, fuzzy systems, ...) to turn our problem into a classification problem (the classes could be : to prune the link or not).

To obtain a complete DB on a full map is clearly not an easy task, because of the large number of links (for instance, 19562 edges in the Argentina map). But the incremental process described in the section 4 might be useful.

References

Type	**Attributes**
Link-based	W_{ij}, F_i, F_j, C_i, C_j, SFL_{ij}, DFL_{ij}, CEN_{ij}, BOR_{ij}
Node-based	F_i, C_i, L_i, L_{SFL_i}, L_{DFL_i}, m_i, M_i, μ_i, m_{SFL_i}, M_{SFL_i}, μ_{SFL_i}, m_{DFL_i}, M_{DFL_i}, μ_{DFL_i}, C_{WL_i}, CEN_i, BOR_i
Global	N_G, L_G, L_{SFL_G}, L_{DFL_G}, m_G, M_G, μ_G, m_{SFL_G}, M_{SFL_G}, μ_{SFL_G}, m_{DFL_G}, M_{DFL_G}, μ_{DFL_G}
Attribute to predict	*need_to_be_pruned* (*true*, *false*)

Table 4: List of available attributes to turn the soft-pruning problem into a classification problem.