The Arab Spring : a highly broadcasted revolution

Robin Leurent *robin.leurent@epfl.ch*
Sacha Leblanc *sacha.leblanc@epfl.ch*
Matthieu Baud *matthieu.baud@epfl.ch*
Alexis Dewaele *alexis.dewaele@epfl.ch*

Abstract

The “Arab Spring” is a historical period dating from December 2010 until mid-2012 that saw the rise of many revolutions and anti-government protests across the Islamic world, starting with Tunisia after years of high unemployment, food inflation, corruption and lack of freedom. Often referenced as the Facebook revolution and as the influence of online media grows more and more every day, it is relevant to study the impact of these media on the spread of revolutions across around 20 countries.

1 Introduction

The purpose of this project is to investigate how the media, social or otherwise, have influenced the happenings of the historical period known as the Arab Spring. To do this, we emit the hypothesis that the coverage of key events (such as riots, destitutions, etc...) are correlated to the start of revolutions of countries that are neighboring Tunisia. It would also be interesting to find out how the coverage done by government-owned media differs from independent media (e.g. blog posts, independent newspapers). We are also interested in studying whether media are consistent with the information they give out, according to media types (e.g. official media, blog posts, forum,...). Indeed by the end of this paper, we will show that it is possible to find when the various protests, riots, etc... happened by studying the media coverage in the countries involved in these events.

2 Dataset Description

The dataset used during this project is the Spinn3r dataset [1]. It contains over 386 million entries from blog posts, news articles, classifieds, forum posts and social media content between January 13th and February 14th. The content includes the syndicated text, its original HTML as found on the web, annotations and metadata (e.g., author information, time of publication and source URL), and boilerplate/chrome extracted content.

Content Type	**Number of elements**
Social Media	231’861’650
Weblog	133’683’918
Mainstream News	14’744’094
Forum	5’734’378
Classified	517’373
Review	32’773
Memetracker	2’473
Total	**386’576’659**

Table 1: Dataset description

3 Description of Main Algorithm

3.1 Data Pre-Processing

The data contained in the Spinn3r dataset is extremely dense and therefore needs to be pre-processed before doing any meaningful work. Indeed, the text in the articles/weblogs/etc... contains HTML tags which need to be removed with the **BeautifulSoup** library [Ric07]. Furthermore, the data contains a field called *language* which is useful for the task at hand. However, some of the entries have a language marked as ”unknown” or are blank which is an important issue since these unknown entries represent around 20% of our data. To solve this issue, the library **langdetect** [Shu10] was helpful in the detection of these unknown languages. In order to focus the data on the research questions, we chose to keep the entries with an European language and with Arabic.

[1] `https://www.icwsm.org/data/`

3.2 Identifying interesting posts

The dataset is huge (1.6TB) and contains millions of entries that emerged on the web in January-February 2011, however most of these entries are not related by any mean to the Arab Spring events. As it is a long and tedious task to work with such a large dataset it was necessary to clear it from all the entries that were unrelated to the events. We thus applied some hand-made topic detection to collect only the event-related entries. It is important to note that this process is made much more difficult than what one may think because of the fact that all the entries are not necessarily in English. That is also why we focused our work on only European languages as some languages are simply not printable or even translatable for processing. This is a very precise task in which we need a good recall while still having high precision in order to discard unnecessary data, therefore the F1-score is a perfect metric for this task.

3.2.1 Latent Semantic Analysis

The first idea to do this topic selection is Latent Semantic Analysis (LSA). LSA is an **unsupervised topic detection** algorithm that uses TF-IDF and matrix factorisation in order to generate a **topic space** in which all the document and words can then be expressed in. To use LSA as a topic detection, one need **reference documents** that one know are related to said topic and then compute the similarity of a given document to the reference document in the ”topic space”, then if similarity is high enough, the document can be considered related to the topic. Our approach of LSA was to add to the ”document corpus” a few Wikipedia pages[2] and web pages that are highly related to the Arab Spring as a positive corpus. The rest of the documents is a negative corpus which is composed of a hundred unrelated Wikipedia pages[3]. We then computed the query vector in the topic space for each entry as follow. Let U,S and V the matrices created by the SVD decomposition of the TF-IDF matrix and Q the entry we want to classify. The query vector Q_v is:

$$Q_v = Q \cdot U \cdot S^{-1} \tag{1}$$

We then computed the cosine similarity between the Arab Spring (positive) corpus and the query vector in the topic space. Unfortunately, the method yielded a high recall but a really low precision due to a high number of false positives, and we had to find another technique.

3.2.2 Keyword-Based Topic Analysis

Even if LSA is a good unsupervised approach for topic detection, we still tackled other possible methods to determine if a document was related to the Arab Spring events. The second solution we turned to was a Keyword-Based Analysis of every entry's document to find if it is related to the Arab Spring events. Indeed, we carefully picked/hand-crafted a total of 69 keywords from the Arab Spring Wikipedia page. Because the data is multi-lingual we translated these keywords into the languages we selected (see section 3.1) and we filtered down all the articles containing strictly more than one of these keywords. The threshold was determined after computing the accuracy, precision, recall and F1-Score on several values and different samples of the data that were hand-labeled for ground-truth. We then averaged the values to get the following table 2.

Metric	>1 keyword	>2 keywords	>3 keywords
Accuracy	0.87	0.85	0.8
Precision	0.7	0.75	0.65
Recall	0.8	0.45	0.2
F1-Score	0.75	0.55	0.3

Table 2: Keyword Threshold Analysis

After applying this filtering, we obtain the language distribution in Figure 1 and the distribution of entries per publisher type as shown in table 3. As we can see, the Arabic language has disappeared. Indeed, it is extremely difficult to extract keyword from languages not using the Latin alphabet: in Arabic, the characters of a word will change depending on its position in the sentence, making it impossible to find our translated keyword. We can't translate Arabic posts in English either because of the limitations of the Google Translation API [4].

3.3 Hyperlink Network Analysis

Let's further our analysis of the dataset by using the hyperlinks contained in the content of differ-

[2] Arab Spring, Egyptian Revolution of 2011, Tunisian Revolution, 2010-12 Algerian protests, 2011-12 Moroccan protests and Yemeni Revolution

[3] Holiday, Sport, Art, IBM, Social media, Weather forecasting, Hank Marino, 2019 Japan-South Korea trade dispute and more

[4] `https://cloud.google.com/translate/quotas`

Content Type	**Number of elements**
Weblogs	698’703
Mainstream News	224’870
Social Media	118’062
Forum	24’891
Classified	840
Memetracker	14
Total	**1’067’380**

Table 3: Distribution of entries per publisher type after filtering

ent websites. In order to create a graph of relations between web pages related to the Arab Spring, we needed to extract the hyperlinks from the content of the entries. To do so, we designed regular expressions to retrieve : the website from the post url, the url links inside the content and finally the websites extracted from the url links. Work on website is useful to cluster the link which is specially interesting for social media post. Indeed the number of Social media post is very large, then clustering their number reduce the graph’s sparsity. The second pre-processing step is to add features to our graph. Therefore we choose to represent the nodes size by their number of outer edges. Besides, the node color depends on the number of days starting from the January 1st or on the post type. Several versions of the graph were drawn with these different parameters (in the notebook). The graph structure (on URL or on website) can be defined as follows:
Each node is a website/url link, and each link is a reference between the two nodes (i.e a relation exists if a website contains the url of the other website in its content). From this we obtain a directed graph (Figure 6).

4 Results and Findings

4.1 Arab Spring Event Retrieval

We’ve seen in the previous subsection, that LSA provides unsatisfying results and from the metrics discussed in section 3.2.2, we’ve seen that Keyword Analysis allows us to filter down only to obtain the entries related to the Arab Spring. With this newly filtered data we can do analysis on event focused entries. For example, verify if we can infer the beginning of the protests or major events in the various countries involved in the Arab Spring by examining the media coverage of these countries. Moreover, we used country specific keywords (the country name, the capital, the leader’s name and the country demonyms), with this we classified each entry as some countries related, e.g. if an entry has the keywords *Egypt* and *Tunisian* we would classify it as Egypt and Tunisia related. From this comes really pertinent graphs (see in appendix B) that show the media coverage per country as a function of time. We selected the most relevant ones (all the other country graphs are in the notebook). As we can see the different figures, in appendix B, we did not manage to infer the beginning of the revolutions for every country. Indeed, in the case of Tunisia (Figure 3), we can see that the red vertical line matches a sharp spike in media coverage, which means we can find when the protests in Tunisia started. In the case of Egypt (Figure 2), we can’t quite infer the beginning of the revolts: the media coverage has a sharp spike a few days after the actual start, but we can predict major events such as the day when President Mubarak resigned, as shown by the green vertical line. It is also the case of Jordanian media coverage: we can’t find the beginning of the protest but we can find major events such as the dismissal of the government by the Jordanian king. However in the case of Djibouti (Figure 5), we cannot infer any events in particular. The limit of event inference can be explained by the fact that we obtain much less data for some events and considering the low amount of information with the noise it becomes really problematic. (like Djibouti’s protests the peak is at around 100 entries in one day compared to a peak at around 7000 entries in one day for Tunisia).

4.1.1 Hyperlink Network Analysis

Since the dataset is huge, the number of entries (even when keeping only event-related ones) is way too large to be plotted in a manner that we can infer and understand something from it. We sampled the dataset for the various graphs but the interpretations can be extended to the full data. One of the first striking analysis is that most of the entries are discarded because they are not containing any hyperlink, also when looking at Figure 6 we see all the nodes lying on the side are very weakly connected. This is very interesting to see that a very high majority of the entries have < 2 links (90% of the entries have no hyperlink at all. And in the remaining 10%, only 3% have more than 1 hyperlink). In the Fig.6 we see the page connections and every node is colored by its type:**black** for Mainstream News, **red** for Forum, **dark purple** for Weblogs, **light purple** for the Social Media. **Yellow-white** nodes are for entries that were **not** initially in the dataset. One striking information to get from this figure is that the majority of the Social Media entries are very weakly connected: 1 or less link. In comparison we see that the Weblogs tend to have a way higher number of hyperlinks (might come from sources or redirections). There are other graphs and analysis in the Notebook.

5 Conclusion

All in all this dataset is very enlightening when we analyse it on a specific topic like we did . It is very interesting to notice that despite having mainly low structured data (76% of the dataset is composed of social Media or Weblog entries and only 21% of it is mainstream news) we are still able to determine very precisely when the key events occur. However the dataset is very sparse in its structure and it can be really challenging to collect information when the topic, the language and even the semantic is different from one entry to another. A final and important point to remark is the unbalanced coverage of the all the events. This is mainly due to the European centered analysis we did, as Europe media will often relay information from the same few countries and show interest for trending news.

References

[Ric07] Leonard Richardson. *Beautiful soup documentation*. 2007.

[Shu10] Nakatani Shuyo. *Language Detection Library for Java*. 2010. URL: `http://code.google.com/p/language-detection/`.

A Language Distribution

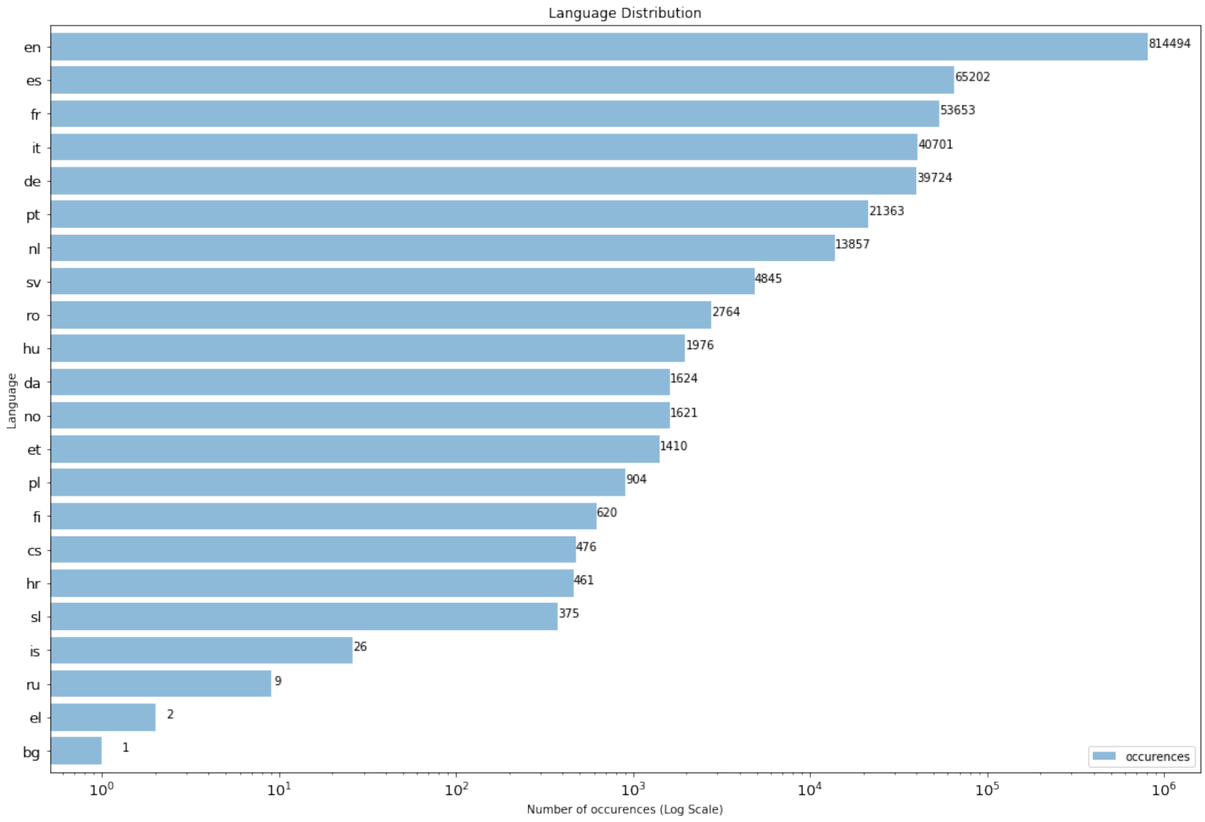

Figure 1: European language distribution of Arab Spring related entries

B Media coverage per country

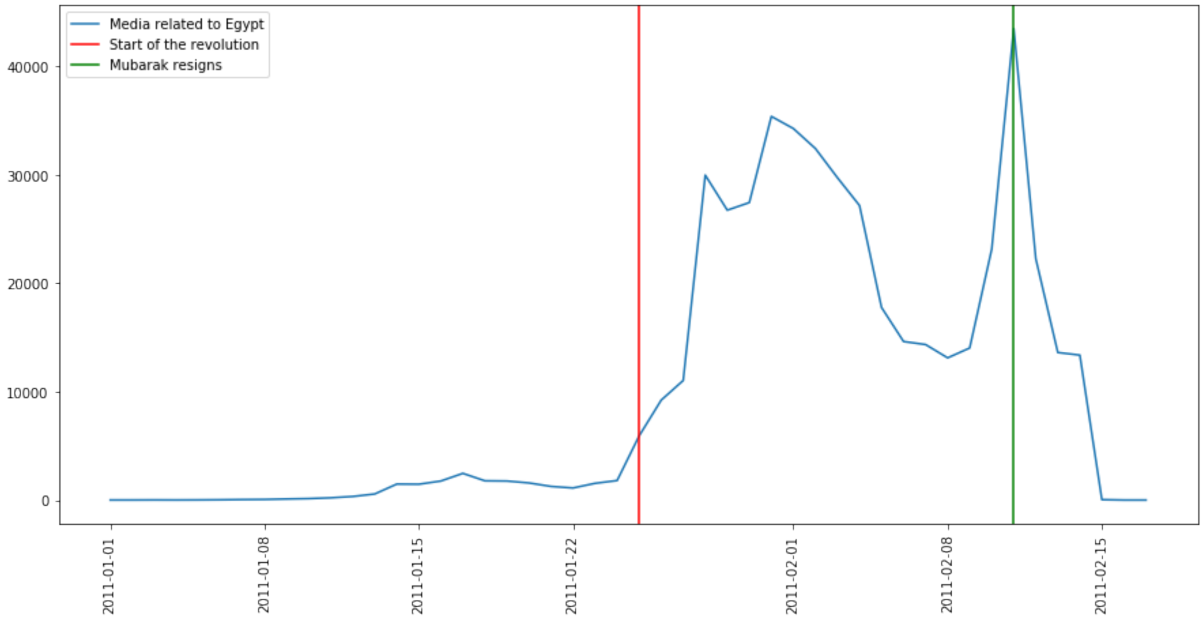

Figure 2: Media coverage in Egypt and main events

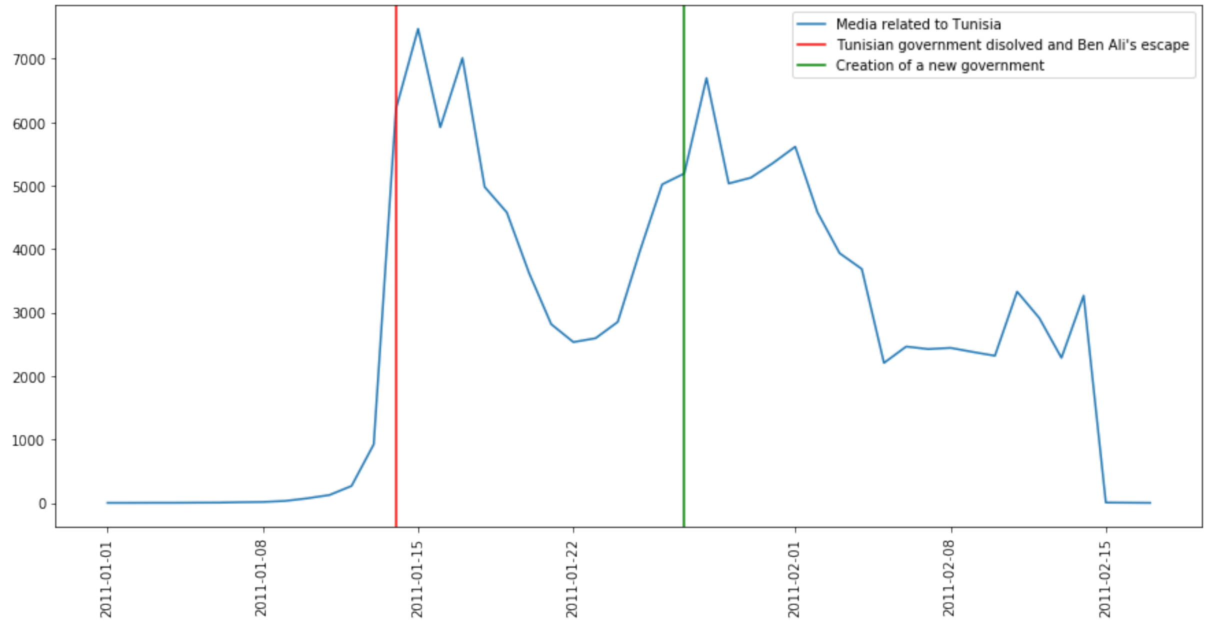

Figure 3: Media coverage in Tunisia and main events

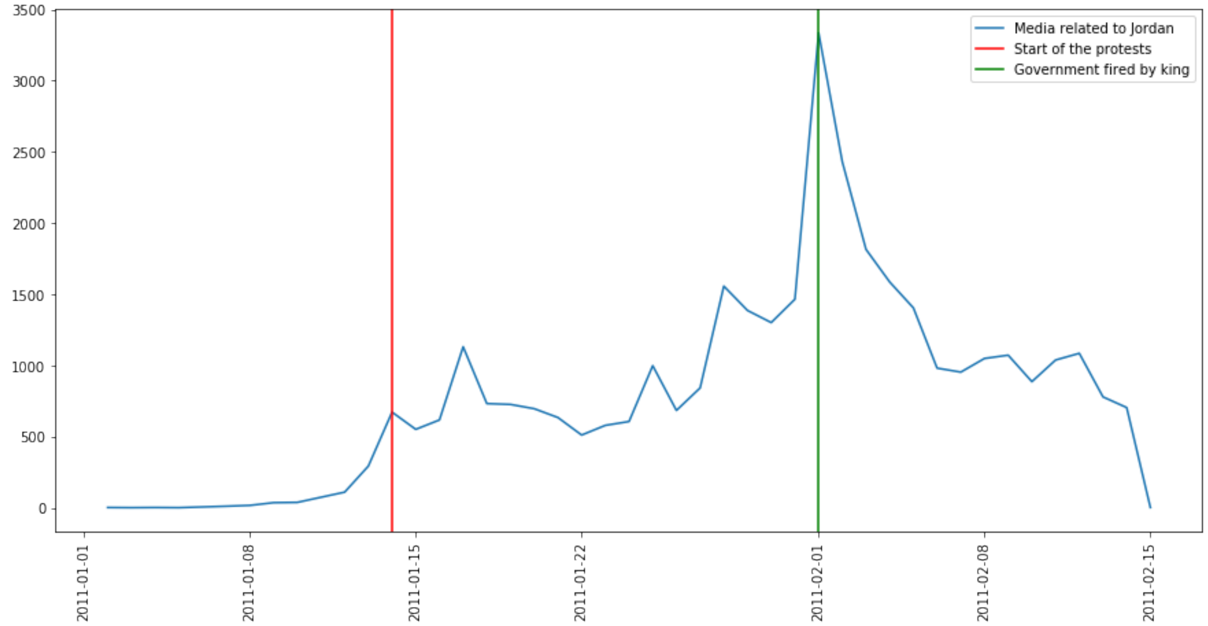

Figure 4: Media coverage in Jordan and main events

C Hyperlink graphs

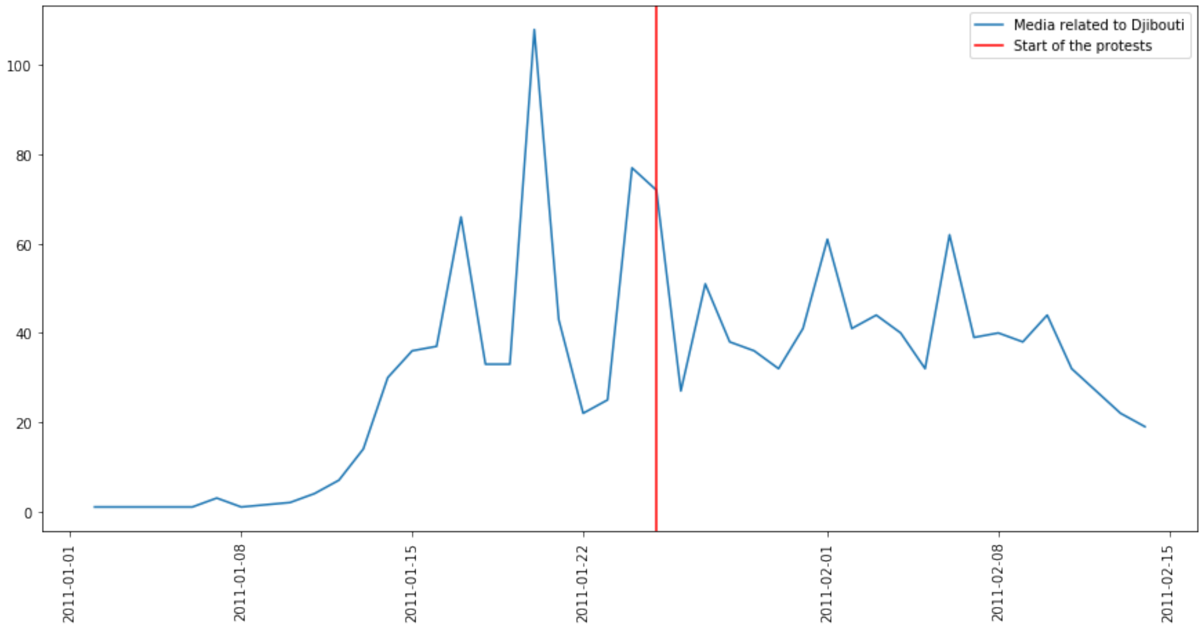

Figure 5: Media coverage in Djibouti and main events

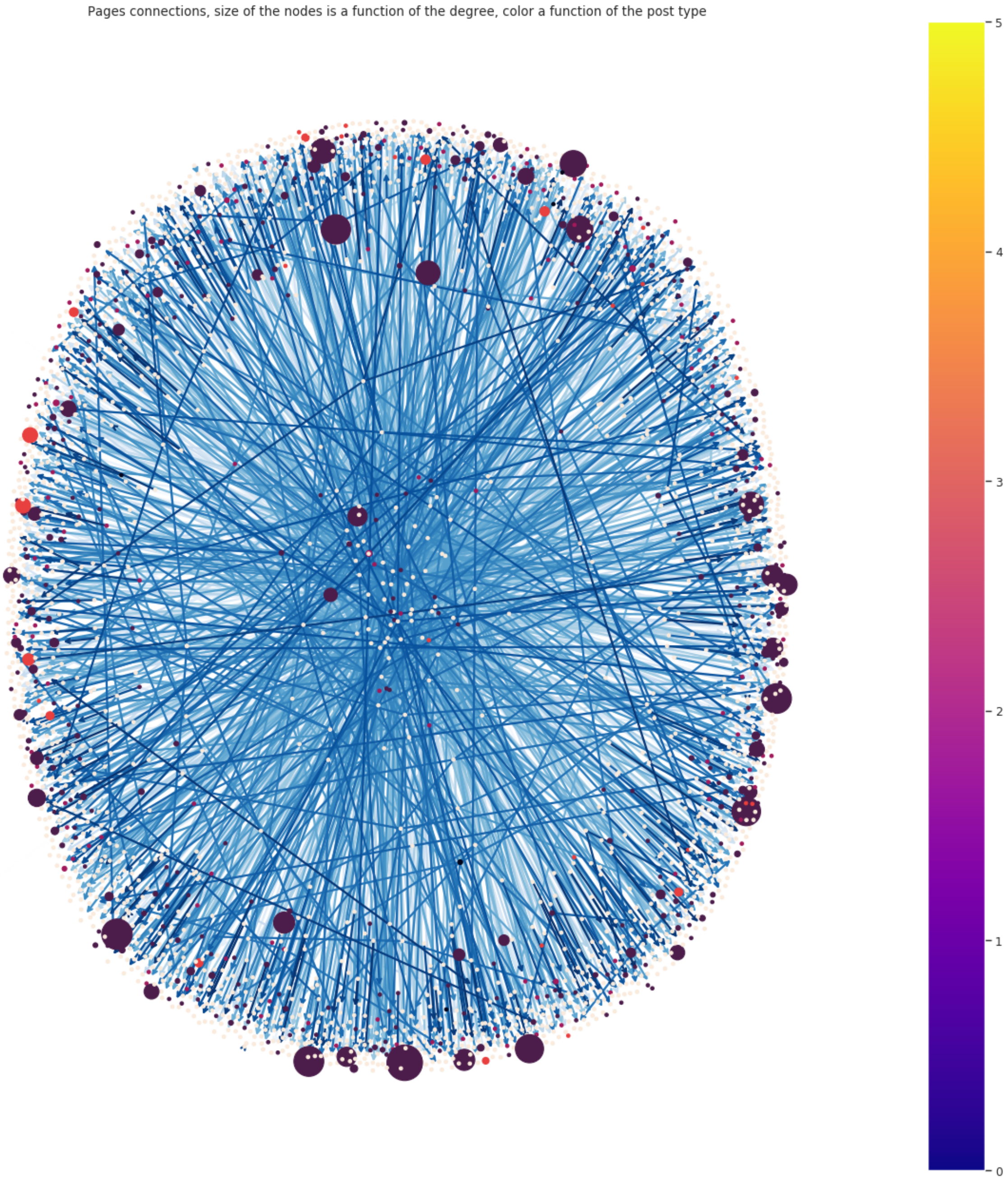

Figure 6: Webpages connections