Université Libre de Bruxelles

Study of the topology of a network of jazz musicians

MEMO-F403

Julien Baudru

Supervised by Hugues Bersini

May 10, 2021

Contents

1 Introduction

There are many systems that take the form of networks, i.e. a set of nodes connected to each other by edges. Among the systems most commonly studied in the literature, we find, among others, the network of hypertext links on the World Wide Web (www), the network of scientific citation [1], the network of roads between cities in a country [2] or the networks related to biology.

This document focuses in detail on the network formed by jazz musicians. In this network, each node represents a musician and the links between these nodes indicate whether they played together on an album or at a concert. These lines will therefore focus on the study of a collaborative network as well as on the construction of this network and its associated parameters.

1.1 Research question

This document is an introduction which attemps to answer the following question : ***What are the parameters favoring the preferential attachment among jazz musicians within a collaborative network ?***

In other words, this paper will try to find out what are the parameters that favor the fact that musicians who have already consecutively collaborated are more likely to make collaborations with new musicians entering the network.

2 State of the art

Regarding the advances in the field of network topology, the latest major discoveries are the scale-free networks and the phenomenon of preferential attachment. These two characteristics, put forward by *A.-L. Barabási* and *R. Albert* in the paper *Emergence of Scaling in Random Networks* [3], have challenged the small-world network models and have also allowed a more general understanding of network topology. The study of scale-free networks has offered the possibility to highlight properties common to many networks in sometimes very distant research fields [4].

The field of natural language processing, which is briefly discussed later in this document, is a particular field of machine learning that focuses on the understanding of texts by algorithms in order to predict a possible classification or translation (or many other purposes). Thus the main goal of natural language processing algorithms is to enable machines to understand and interpret human speech and text. The first creation of modern NLP algorithms seems to date back to the end of the 1980s thanks to a mix of linguistic and statistical methods. Today, there are several models that have been considered for a while as being the most performing, we can mention among others the models of Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks in particular. In 2017, researchers from *Google*, in their paper named *Attention Is All You Need* [5], suggested a model called the *Transformer* that currently seems to be the most successful in the field of natural language processing.

Regarding the collaboration networks of jazz players in particular, existing research is not extensive. However *P. M. Gleiser* and *L. Danon*, in their paper entitled *Community structure in Jazz* [6], have put forward the primordial parameter of communities, i.e. the grouping of nodes into sets supposed to share common characteristics. One of the first algorithms to detect communities within a network was based on centrality indices to find the boundaries of different communities and was suggested by *Newman* and *Girvan*[7]. In their article *Community structure in social and biological networks* [7], they provided a new method with a high degree of success for identifying communities in networks. The concept of community is particularly important and relevant for networks linking humans, as is the case for the jazz musician collaboration network. Based on the method suggested by *Girvan* and *Newman*, *Gleiser* and *Danon* have highlighted the presence of racial communities in the jazz world between 1912 and 1940. At that time, most of the white musicians performed only with other white musicians and the same was true for black musicians. In addition, most bands were exclusively comprised of one race or the other. Nowadays, based on the concept of modularity which is a quality index for a partition of a network into communities, the *Leuven* algorithm is the one with the best global modularity. However, as *Fortunato* and *Barthélemy* have shown [8], it seems important to specify that the best modularity is not synonymous with best communities. It should be noted that community detection algorithms can be divided into two families, the static and the dynamic ones. These two families can be further subdivided into two where we find the algorithms allowing the overlapping of communities (a node can belong to more than one community) and the algorithms that do not allow it. In this document, concerning the communities, we will only talk about static algorithms that do not allow overlapping. However, regardless of the algorithm chosen,

it is still difficult to interpret the communities produced by these algorithms without the help of additional information.

Finally, a last area discussed in this paper is the preferential attachment. The study of this phenomenon seeks to determine what factors influence the creation of new links during the dynamic evolution of a social network. The concept of preferential attachment was introduced by *Barabasi* and *Albert* [3], who showed that nodes with higher degrees tend to attract new nodes (and thus links) during the evolution of the network. Historically, it seems that *Udny Yule* was one of the first to put forward this phenomenon to explain the power law distribution, which is why it is also called the *Yule process*. This process generates a so-called *long-tailed* distribution following a Pareto distribution or a power law. The phenomenon of preferential attachment could be roughly summarized by the sentence: *The rich get richer*. It is more globally known as the *Matthew effect* or cumulative advantage process. The most famous model allowing to simulate the phenomenon of preferential attachment is suggested in 1999 by *Barabasi* and *Albert*: it is based on previous works of the physicist *Derek J. de Solla Price* and his *Price*'s model. In their model, each new node added to the network is connected to existing nodes with a probability proportional to the number of links that the existing nodes already have. In 2009, based on the preferential attachment (BA) model proposed by *Barabasi* and *Albert*, *E. Ben-Naim* and *P. L. Krapivsky* showed that in a preferential attachment network the degree distribution of the nodes depends on the depth (the latter being defined as the distance of the node from the root of the network) [9]. Moreover, they showed that nodes closer to the root tended to have a larger number of connections. They explain this phenomenon by the correlation that exists between the depth of a node and its age, so that younger nodes, which are further from the root because they arrived later, are the least connected. In a paper published in *Nature* [10], *A. Topirceanu*, *M. Udrescu* and *R. Marculescu* showed that the degree of the node is not the main attractor of new social links. They showed that the betweenness of the nodes and the strength of the links play a crucial role in the preferential attachment and thanks to that they suggest a new model named *Weighted Betweenness Preferential Attachment* (WBPA) model. In 2003 [11], *H. Jeong*, *Z. Néda* and *A.L. Barabási* showed with a method to quantify preferential attachment in evolutionary networks that this phenomenon was indeed present in real networks. In their studies on four networks (scientific citation network, internet, actor collaboration and science coauthorship), they found that the rate $\Pi(k)$ with which a node with k links acquires new links is a monotonically increasing function of k either linear, power-law, or sublinear. More recently, among the new models proposed, *S. Ruj* and *A. Pal* are the first to put forward a preferential attachment model with degree bound [12]. In this model, the maximum degree is upper bounded by a fixed value and according to the authors, this model is more suitable for IoT and cyber-physical systems than the conventional preferential attachment model.

3 Methodology

In order to answer the question stated in point 1.1, the research work is divided into two main parts: the construction of data sets (4) and the study of the multiple parameters of the network (5). The first part, the data collection, uses artificial intelligence algorithms of type NLP (Natural Language Processing) and web scrapping to extract useful information from web pages used as sources. Once these data are collected, the static and dynamic networks can be built. Then in the second part, the different parameters usually studied in the networks are analyzed and explained.

4 Data collection and network building

4.1 Building data sets

Each dataset is composed of 4 columns: The name of the album/festival, the release year of the album/year of the live performance, the label (album only) and the list of musicians present on the album/live performance.

The list below compiles the sources that allowed (via web scrapping) to build the datasets:

1. Jazzwise: The 100 jazz albums thaht shook the world for *data.csv*
2. Wikipedia: Liste des albums de jazz les plus vendus for *data1.csv*
3. Wikipedia: Album de jazz for *data2.csv*
4. Wikipedia: Album de jazz Américain for *data3.csv*
5. Wikipedia: Album de jazz français for *data4.csv*

6. Wikipedia: Album de jazz fusion for *data5.csv*
7. Wikipedia: Album de bossa nova for *data6.csv*

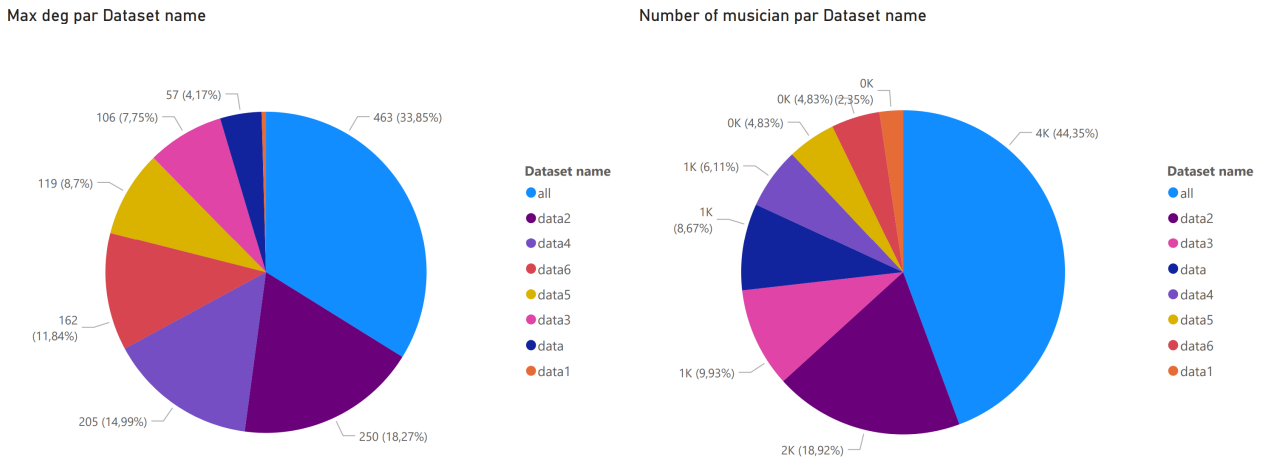

Figure 1: Data sets comparison

The graphs presented in the figure above highlight the fundamental differences in the data sets used to build the network. At the beginning of the project, a first data set simply named *data.csv* was built by hand. Then, the dataset named *data1.csv* was built using simple web scrapping methods and the Pandas (Python) library. We notice that the number of musicians in this dataset is much lower than in the other datasets. This is due to the difficulty of retrieving only the names of the musicians using only simple web scrapping methods. Finally, to overcome the problem encountered in the *data1.csv* dataset, all other datasets were built using NLP (Natural Language Processing) algorithms in combination with web scrapping (e.g.: *data2.csv*, *data3.csv*, etc.). Information on the operation and use of these particular algorithms is detailed in the section 4.2 of this document.

In addition to the datasets associating the album/live performance with the list of participating musicians, other data were collected during the different research phases. These data included the birthdate, the birthplace, the labels and the favorite musical sub-genre of the musicians. The dates of the collaboration in the main date sets are used for the construction of the dynamic network (section 4.3.2), and gender and ethnicity were abandoned for reasons explained in section 4.2 .

4.1.1 Information on data sets

Once the different data sets are combined, the database includes a total of 3,602 different musicians and nearly 400 different albums and live performances. These data span a period of time from around the year 1928 to the present day. As expected, the data collection is incomplete given the huge volume of albums and live performances that are added each year in addition to the many albums that are not cited by the various websites from which the data comes. Like many social networks, the collaborative network of jazz musicians is a growing network. Moreover, ∼4,000 nodes is a relatively small number compared to the size of networks usually studied in the field, as a comparison the Timik platform network treated in the following document [10] counts nearly 364,000 nodes. Besides, it seems important to note that six out of seven sources are from the French-speaking community of *Wikipedia*.

It seems important to point out that since the datasets are built with information provided on *Wikipedia*, which is a collaborative website, they may be subject to inaccuracies and gaps. Another essential parameter to take into account is the fact that these datasets are constituted via webscrapping methods detecting all the names of people (or looking like them) present on the Wikipedia page of an album/concert. Thus, although it is rarely the case, some articles quote musicians who did not collaborate on the album; when they are quoted, it is often for comparison or anecdotal purposes. This has as a direct consequence: some links are made between musicians when they did not really collaborate. Fortunately, the effects of this problem seem negligible at the global network level and the datasets give an acceptable overview of the jazz scene. In addition, the filters introduced in the next section allow to reduce its side effects considerably.

4.2 Usage of natural language processing algorithms

As briefly explained above, the main function of the NLP algorithms in this project is to locate the different names of the artists present on a web page of an album or a live performance. Before using the NLP methods, a web scrapping algorithm, which takes place in two steps, is set up. First, based on a Wikipedia page listing albums/concerts from a certain country, style or other criteria, the algorithm will fetch all the links leading to the pages of the different albums/concerts mentioned in this main page. Then, for each of these links, so each album, the algorithm will fecth the HTML code of the page pointed by this link. To briefly summarize the data collection, the NLP algorithm related to webscrapping works as follows:

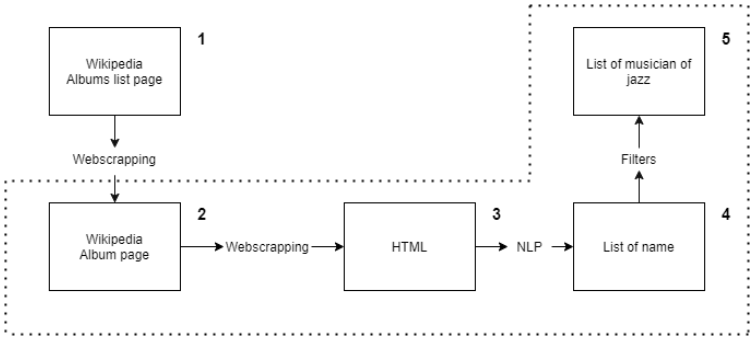

Figure 2: Steps for getting the jazz name

The NLP algorithm extracts from the HTML page the sets of words that are recognized as being of type *PERSON*. This type is provided natively in the *nltk* library. The common names that can be erroneously recognized as *PERSON* by the NLP algorithm are filtered using the following word lists provided by *nltk*:

```
set(nltk.corpus.words.words('en'))
set(nltk.corpus.words.words('fr'))
```

Moreover, others filters are applied to the collected data. These filters avoid counting producers, authors, journalists, labels and album names often cited in the *Wikipedia* articles who are sometimes recognized as false positives by the NLP algorithm. Depending on the category, each filter contains, for example, a list of the most frequently cited jazz producers in a simple text file. Of course, the professions of producer, lyricist and journalist are closely related to music and are of considerable importance in the jazz scene. However, they do not fit into the question that this document tries to answer, so it was decided to leave them out. Thus, the purpose of these filters is to reduce the noise present in the collected data.

Another possible use of natural language processing algorithms would be to recognize when several different names refer to the same musician. For example, the pseudonym of the jazz composer *George Gershwin* is *Jacob Gershowitz*. We often notice the presence of pseudonyms for musicians too. Most famous pseudonyms were managed by hand.

Moreover, another machine learning application initially present in this project used the ethnicolr library based on *Tensor Flow*. The purpose of this algorithm was to add the ethnic origin of the musicians in the datasets in order to compare the results obtained by *Gleiser* and *Danon* about the racial segregation between 1920 and 1940 with the data used in this project. The aim was also to eventually show an evolution of mentalities after 1940. The choice of using machine learning for this task was made because data on the origins of musicians are rarely explicitly provided on the Wikipedia pages used as sources and can therefore not be obtained via classical webscrapping. More importantly, given the large number of musicians present in the datasets, this data collection seemed too tedious to be done by hand. To do so, the algorithm tried to guess the origin based on the names of the musicians. The resulting classes were: Non-Hispanic Whites, Non-Hispanic Blacks, Asians, and Hispanics. But since this library is based primarily on consensus from 2000 or 2010 in the United States of America, the origins of jazz players prior to this period are often awkwardly predicted. Moreover, in the same way as for the Belgian rapper network (discussed in the next section), a lot of jazz musicians use pseudonyms, which deeply complicates the detection of the origin of the musicians for the algorithm. The use of pseudonyms also causes difficulties in detecting the gender of musicians. The gender-guesser library in charge of this task is based on statistics from a list of 50,000 names. In various tests this librairy has often tended to identify the gender of most musicians as *unknown*.

Furthermore, race, ethnicity and gender are complex and controversial topics, requiring knowledge that the author of this document does not possess. For this reason, this part of the data collection has been abandoned.

4.3 Building the network

The section presented here details, on the one hand, the construction of the static network and, on the other hand, the construction of the dynamic network.

4.3.1 Static network

In the studied network, every link between two nodes has a weight. This weight is simply determined by the number of times two musicians have collaborated during a recording or a live performance. In addition, the network constructed here does not have any directed edges. Regarding the nodes, those whose names have been identified as male (or close to it) are drawn in orange while those whose names have been identified as female (or close to it) are drawn in green. But as explained in the previous section, the identification of genera is not perfect for the studied network.

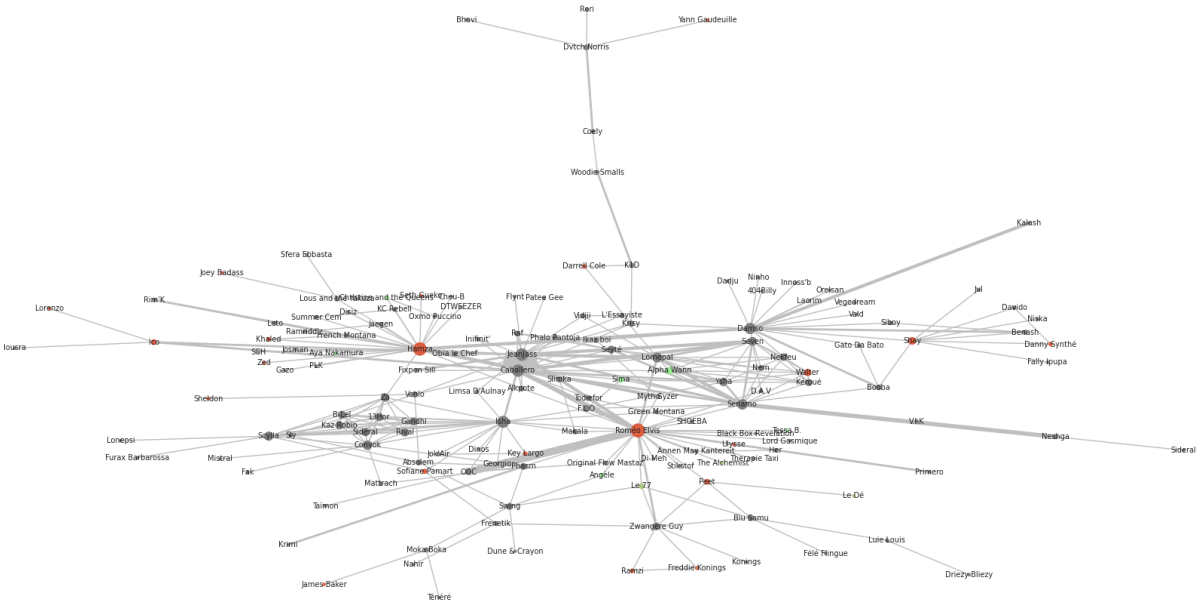

Figure 3: Belgian rap collaboration network

The figure above represents the collaboration network for the release of a single for Belgian rappers (what is more commonly called featurings). This example of a network, close to the one studied, was chosen because of its smaller size. This reduced size allows the reader to more clearly identify the weight differences the links on the graph.

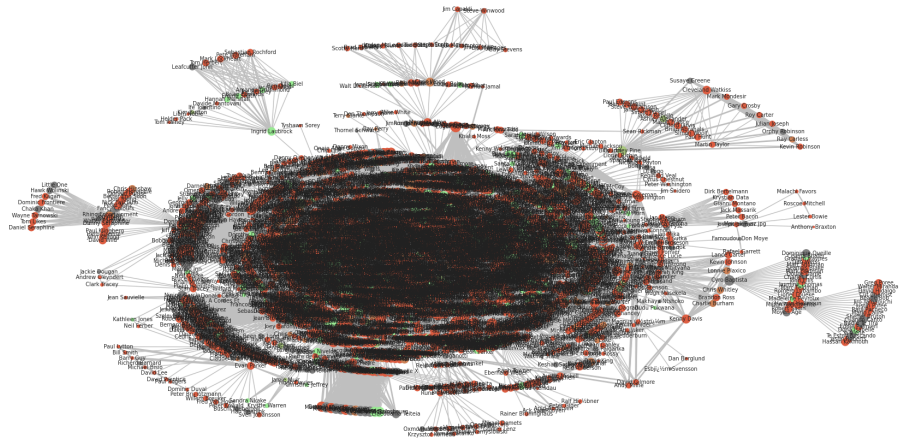

Figure 4: Jazz collaboration network with a kamada kawai layout

The figure above is the network resulting from the combination of the different jazz data sets mentioned before. The shape taken by this graph is due to the *Networkx* layout function named *kamada_kawai_layout*.

Force-driven graph drawing algorithms is a class of algorithms used to create graphs in an aesthetic way. This same network can be represented in a circular way as below, but the circular formatting of the network makes the reading almost impossible.

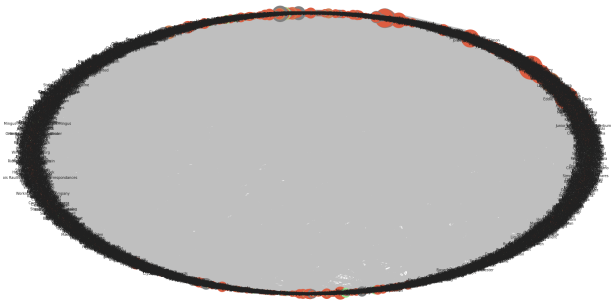

Figure 5: Jazz collaboration network with a circular layout

The graphical construction of the network must be optimized; for the moment, it uses the *Networkx* and *Matplotlib* libraries. The slowest part of the process is, as expected, the creation of the numerous links between the nodes of the network as well as the drawing of the network. Some libraries such as *neo4j*, *graphistry*, *Cytoscape3* or *BioFabric* seem to be suitable solutions for a more efficient implementation, although given the small number of nodes in the network studied here (∼4,000) it is not a priority.

4.3.2 Dynamic network

In order to study the preferential attachment phenomena that occur in the network, a dynamic network was developed. Indeed, the study of the preferential attachment requires adding a time variable to the network in order to see the evolution of the links between the nodes constituting it. The fundamental difference with the method presented above lies in the creation of the links. Each link, without weight in this case, when it is created, contains information about the time at which it appeared in the network. In the context presented here, this information is the album release date (or the date of the live performance). Since *Networkx* does not allow to add other information on the links than the weight, the DyNetx library was used to add the collaboration dates between musicians.

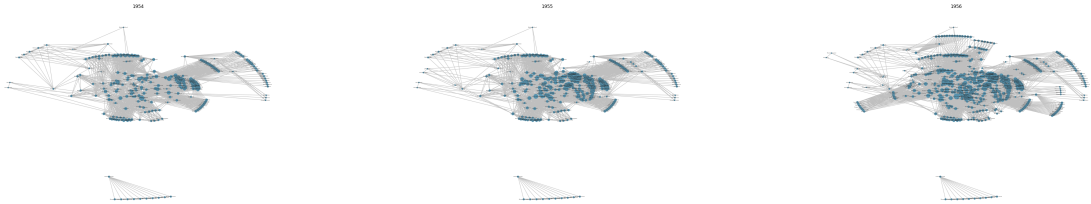

Figure 6: Network evolution sample

In practice, to allow a better visualization of the network evolution of the network, a video from the overlay of the subnetworks of each year has been created. You will find the final result here.

Note that the dates used to describe the links between the different nodes are obtained thanks to a similar web scrapping method presented in section 4.2. As a consequence, some nodes are not present in the dynamic network because no information was available regarding the release date of the album/performance on the web pages used as sources.

5 Experimental analysis

5.1 Parameters studied

The different parameters studied in this section provide an overview of the network. These parameters are those usually studied in other documents dealing with collaborative networks, such as the rich-club and clustering coefficients, the gamma factor, and the notion of communities.

5.1.1 Factor γ and scale-free networks

One of the essential parameters of scale-free networks is their power-law distribution. In these networks, the degree distribution of the nodes follows a power law. Thus, there is a high occurrence of low degree nodes and a low number of high degree nodes. The gamma factor is used to determine if a network is scale-free, which is calculated via the following equation:

$$P(k) \sim ck^{-\gamma}$$

where *P(k)* is the frequency such that $P(k) = \frac{n_k}{n}$, c is a proportionality constant and k is the degree. A network is generally determined to be scale-free if the γ value is between 2 and 3. *P(k)* can also be interpreted as the probability that a node is k links.

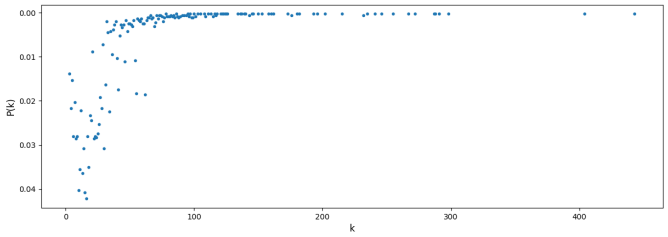

Figure 7: Frequency of node degrees

The value of the γ factor for the studied network is 2.048. This value of γ was found thanks to the curve fitting method *Fit* of the powerlaw library. Thus, it appears that the network corresponds to the characteristics expected for a scale-free network.

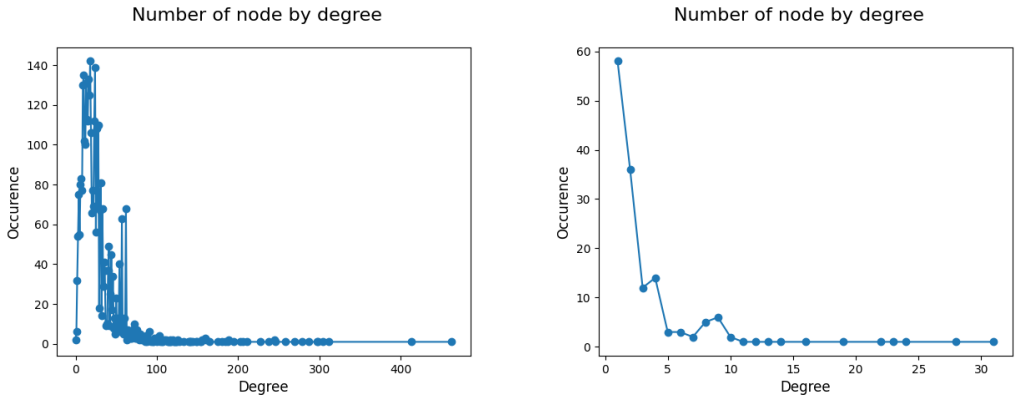

Figure 8: Degree distributions

Regarding the graphs above, the one on the left represents the distribution degrees of the nodes of all the combined datasets. The one on the right represents the distribution of the dataset *belgian_rap.csv* cited in the section 4.3. In both cases, it appears that there is a large number of low degree nodes and a small number of high degree nodes. The average degree of the studied network is 91.28 while the average degree of the Belgian rapper network is 12.16.

5.1.2 Rich-club coefficient

The rich-club coefficient allows us to check whether vertices of high degree tend to be strongly connected to each other. This coefficient is calculated via the following equation:

$$\phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k} - 1)}$$

where $E_{>k}$ is the number of links among the $N_{>k}$ nodes of degree higher than k and $N_{>k}(N_{>k} - 1)$ is the maximum number of links between the $N_{>k}$ nodes. This coefficient is normalized using the rich-club coefficient of a random graph of the same order as the one studied. The normalized indicator is therefore the following:

$$\rho ran(k) = \frac{\phi(k)}{\phi ran(k)}$$

For this metric, if for certain values of k we have $\rho ran(k) > 1$, this denotes the presence of the rich-club effect.

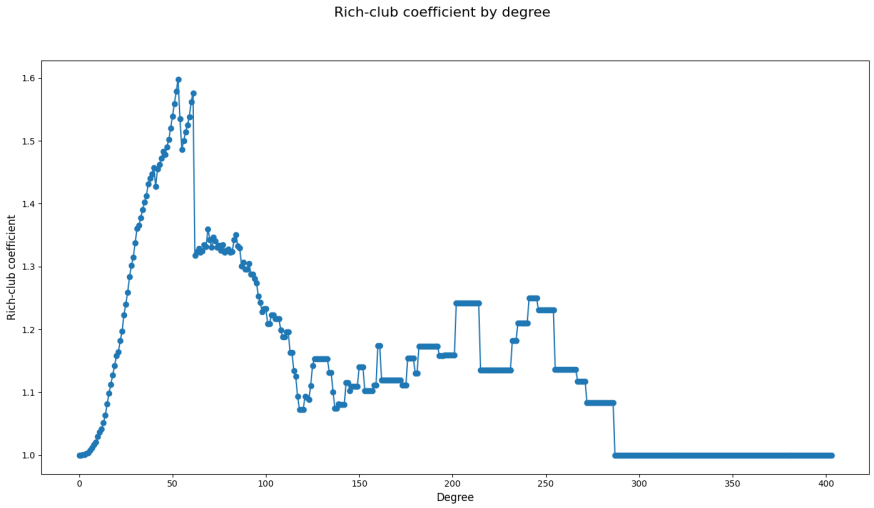

Figure 9: Distribution of rich-club coefficient by degree

Thus, in this network, the nodes for which the rich-club coefficient is the highest are those of degree k = 61, their rich-club coefficient being 1.58.

The rich-club coefficient is calculated via the *rich_club_coefficient()* method of the *NetworkX* library, the average value of this one for the complete network being 1.23. It can be noticed that for nodes of degree k close to k_max (443), the rich-club coefficient is close to 1, which means that the nodes of high degree are well connected to each other. In a network different from the one studied here (for example: Internet), this would mean that the network is robust and that removing a hub would not affect the general connectivity of the network.

5.1.3 Modularity and community

Most social networks highlight the notion of community [7] bringing together the different members, called nodes in these pages, of the network. There are different methods to detect communities within a network such as the method of *Clauset*, *Newman* and *Moore*, the method of *Pons* and *Latapy*, the method of *Watike* and *Tsurmi* or the method of *Louvain*. The one used in this document is called the *Louvain* method [13] created by *V. D. Bondel*. This method has been chosen because it is easily implemented thanks to the python-louvain library and especially because it seems to be the most efficient method to date.

Modularity Optimization Comparison[9]

	Karate	Arxiv	Internet	Web nd.edu	Phone	Web uk-2005	Web WebBase 2001
Nodes/links	34/77	9k/24k	70k/351k	325k/1M	2.6M/6.3M	39M/783M	118M/1B
Clauset, Newman, and Moore	.38/0s	.772/3.6s	.692/799s	.927/5034s	-/-	-/-	-/-
Pons and Latapy	.42/0s	.757/3.3s	.729/575s	.895/6666s	-/-	-/-	-/-
Wakita and Tsurumi	.42/0s	.761/0.7s	.667/62s	.898/248s	.56/464s	-/-	-/-
Louvain Method	.42/0s	.813/0s	.781/1s	.935/3s	.769/134s	.979/738s	.984/152mn

Figure 10: Wikipedia: Modularity optimization comparison

This method allows to perform the partitioning of a network by optimizing the modularity. The modularity is a value between -0.5 and 1 which measures the density of edges inside the communities compared to the density of edges connecting the communities. The formula to calculate the density is the following:

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j)$$

where A_{ij} gives the weight of the edge between the nodes i and j; k_i and k_j are the sum of the weights of the edges linked to nodes i and j; $2m$ is the sum of all the weights of the edges of the graph; c_i and c_j are the communities and δ is Kronecker delta function (δ (x,y)=1 if x=y, 0 otherwise).

The communities obtained for the studied network are as follows:

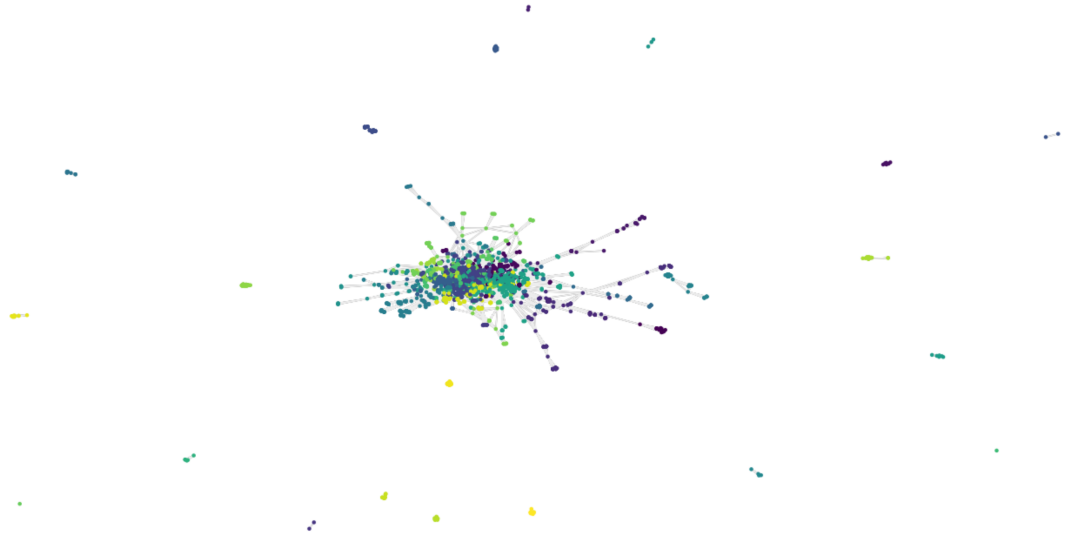

Figure 11: Communities in jazz collaboration network

The Leuven method has enabled to put forward from 47 to 52 different communities in the network of collaboration between jazz musicians. These multiple communities are highlighted here with different colors. The number of communities can vary because the Louvain [2] algorithm is unstable. Indeed, the placement of nodes in the different communities depends, among other things, on the evaluation order of the nodes [14]. Below is the same algorithm applied to the smaller network mentioned above (this time we get ∼10 different communities).

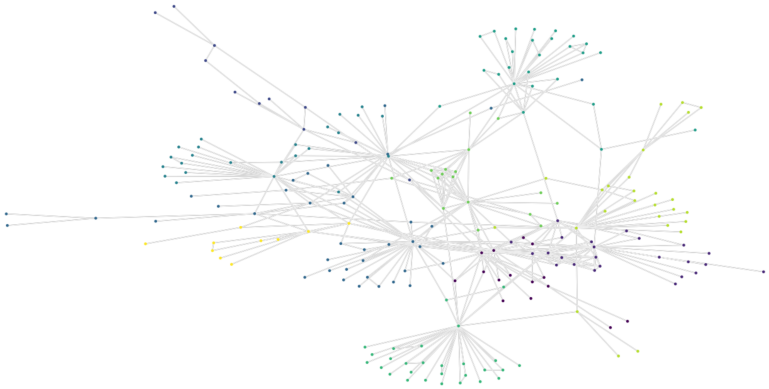

Figure 12: Communities in Belgian rap collaboration network

As mentioned in section 2, it is often difficult to draw conclusions from the communities highlighted by the algorithm without having additional information available. However, in the collaborative network of Belgian rappers, it seems that most communities are built around the important nodes of the network. For this network, we also notice that the communities are divided between different categories of popularity with, on the one hand, communities of mainstream artists and, on the other hand, artists rather categorized as underground. Regarding the collaboration network between jazz musicians, as for the network of rappers, we notice that most of the communities are built around a very connected node. However, as there are clusters of nodes poorly connected to the main network, it appears that sets of nodes logically form communities without the presence of highly connected nodes within them. Moreover, unlike the collaboration network of Belgian rappers which only took into account one country, the jazz musician's network covers many. Thus, it appears that there is a link between the communities and the geographical origin of the musicians, for example one notices the presence of several communities comprising exclusively French musicians. The same is true for the United States of America, Brazil and most of the countries mentioned at section 5.3.2 of this document. When there are several communities for one country, it seems that the difference between these communities is determined by the age of the musicians and by the sub-genre of jazz they usually play.

5.1.4 Clustering coefficient

As defined by *Girvan* and *Newman*, the clustering, or network transitivity, is the property that two vertices which are both neighbors of the same third vertex have a heightened probability of also being neighbors of one another. The clustering coefficient is defined as follows:

$$C = \frac{3 * (\text{number of triangles on the graph})}{(\text{number of connected tropes of vertices})}$$

The coefficient C is the probability that two nodes are connected knowing that they have a neighbor in common. In other words, this coefficient indicates the probability that two musicians having collaborated with a musician on a project have themselves collaborated on another project.

The distribution of the clustering coefficients for the studied network is given by the figure below:

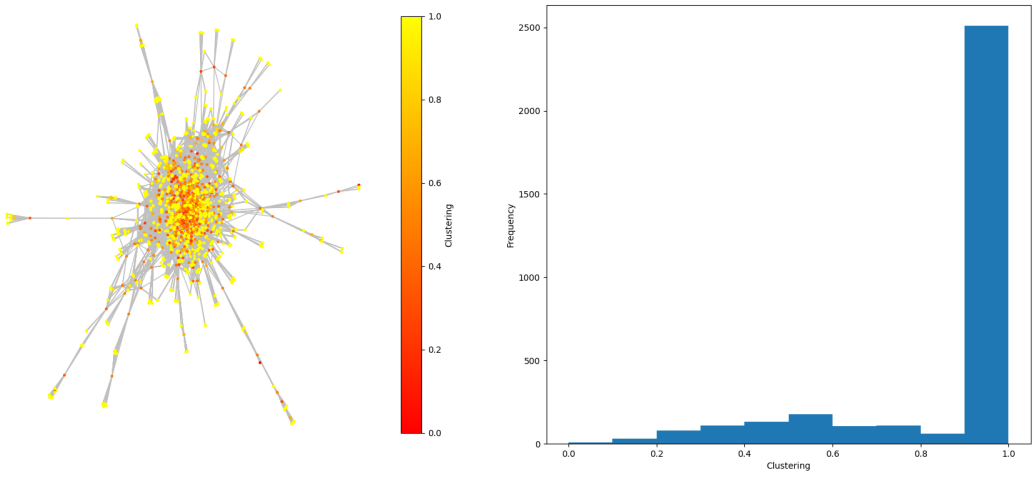

Figure 13: Distribution of clustering coefficient

It appears that most of the nodes ($\sim$ 2,500) in the network have a clustering coefficient between 0.9 and 1. Moreover, the value of the average clustering coefficient is 0.88 and the transitivity value of the network is 0.48. Thus, from the definition of the clustering coefficient, it appears that more than 90% of the nodes in the network have a high probability of being connected to another node knowing that they have a neighbor in common. We can conclude that the network studied here is strongly aggregated.

5.2 Preferential attachment

As explained in section 4.3.2, the study of preferential attachment induces an evolution over time of the network, which is commonly called a dynamic network.

The *preferential_attachment* method of the *NetworkX* library allows to compute the preferential attachment score between two nodes u and v. This score is calculated according to the following formula: $|\Gamma(u)||\Gamma(v)|$ where $\Gamma(u)$ gives the set of neighbors of u. The higher this score is, the more strongly the two nodes u and v will be linked. The graph below shows the evolution of the maximum score of preferential attachment within the network. As expected, it increases as the network grows over the years.

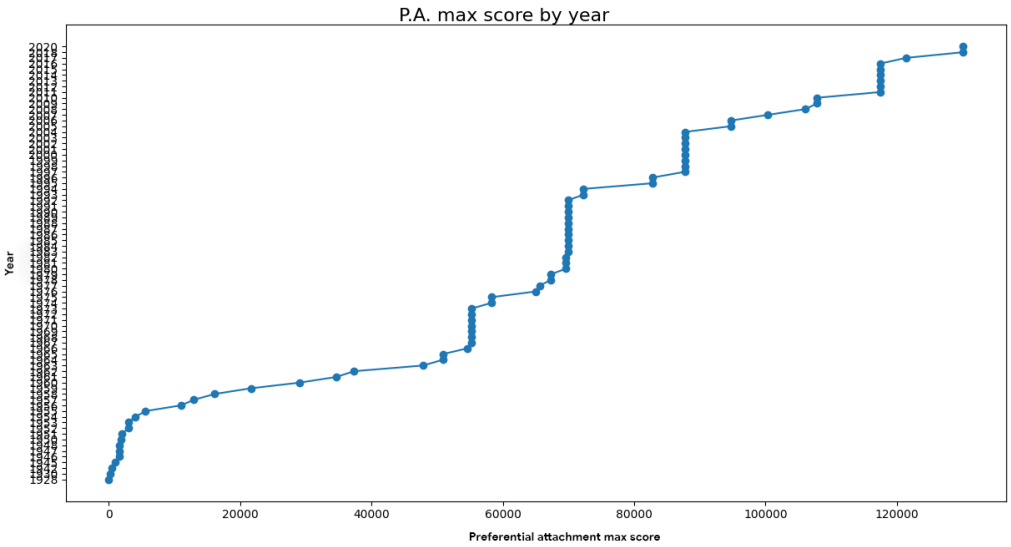

Figure 14: Maximum P.A. score over the years

It can be noted that the pair of nodes for which the preferential attachment score is the highest, for the final network, is the pair of musician *Stan Getz* and *Duke Ellington* with a score of 130103. It appears that these two nodes are both among the top hubs of the network studied, *Duke Ellington* being the most connected node and *Stan Getz* being the fifth most connected node.

It also appears that on four occasions, the maximum preferential attachment score does not vary for several years in a row. This is the case between 1966 and 1973, between 1983 and 1992, between 1998 and 2004 and between 2012 and 2018. This means that new nodes that are added to the network during this period do not attach to nodes with the current maximum P.A. score. It seems important to specify that in this case, the number of new nodes added each year to the network is not constant. Indeed, this number depends on the data available in the datasets, so between 1928 and 2020, not all years are present. According to the data collected, there are no new nodes for the following years: 1929, 1930 to 1942, 1943 to 1945, 1949 and 2019.

5.3 Analysis of the information carried by the highest degree nodes

In an attempt to draw conclusions from the information carried by the 50 nodes with the highest degree in the network, these different parameters will be compared to the information carried by 50 randomly selected whose degree is between the first quartile (Q1) and the second quartile (Q2). The sample size was set at 50 because the collection of some data for these musicians cannot be automated (for example: place of birth, labels, etc...). In addition, it also seems important to specify that due to their smaller popularity (on the Internet), some data concerning the sample of musicians whose degree is between Q1 and Q2 are more difficult to obtain.

5.3.1 Instruments

The graph below shows the 50 most connected nodes of the network with their preferred instrument. The degrees of the different nodes are also shown on this graph. It appears that the two most connected musicians of the network are *Duke Ellington* and *Miles Davis* with respectively degrees of 443 and 404. Moreover, a non-negligible difference of 106 degrees is present between the second and the third biggest nodes of the network which is *Bill Evans* (deg. 298). Then, the following degrees in the ranking decrease gradually, until the last of the 50 most connected nodes being *Jon Hendricks* with a degree of 115.

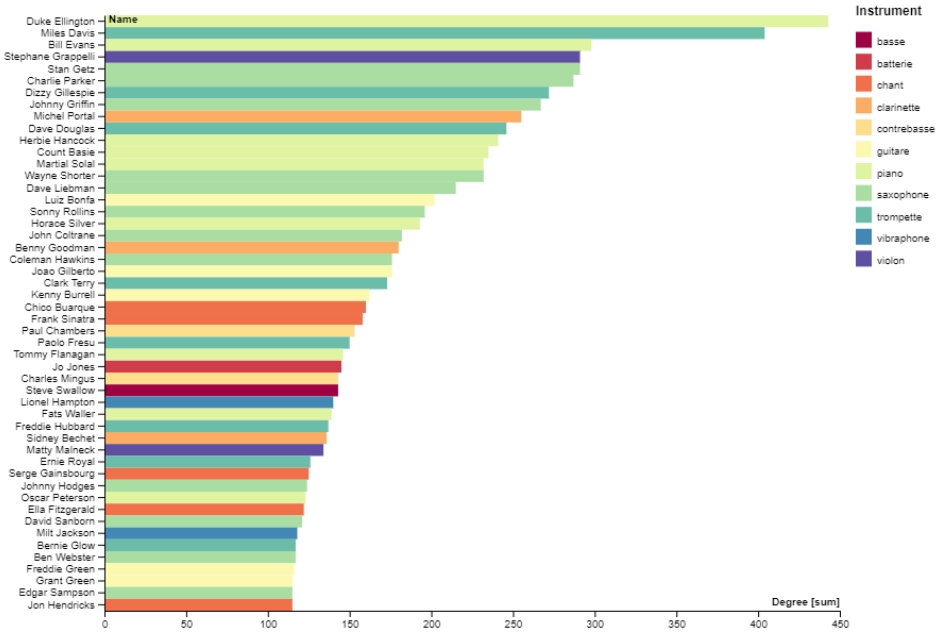

Figure 15: Top 50 nodes with their preferred instrument

In the studied sample, as expected from the power-law distribution, the disparity of the degrees of the nodes is much less marked than in the first sample. The average degree of this set is 12.24. The lowest degree is 11 while the highest is 13.

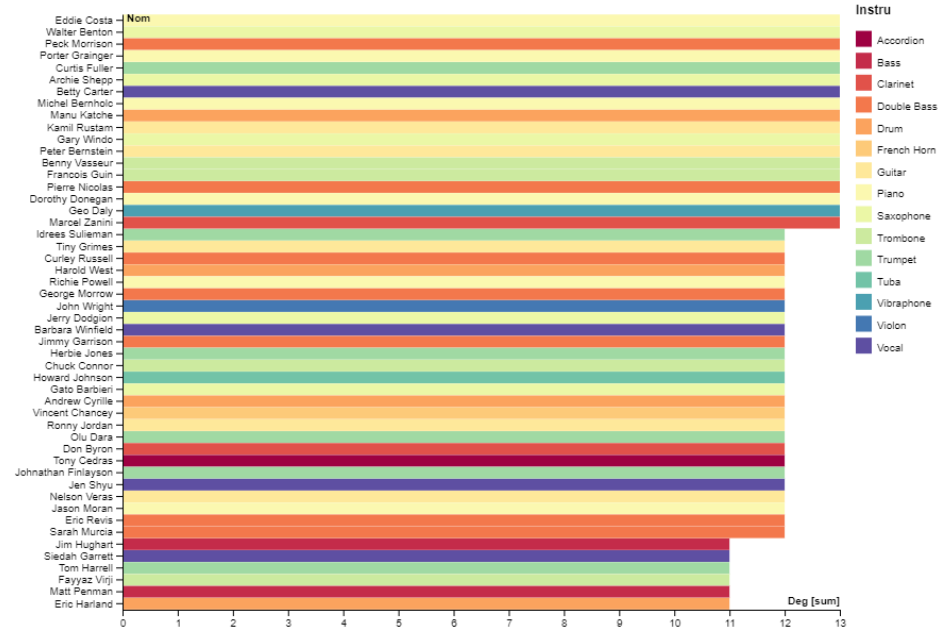

Figure 16: The 50 randomly selected nodes between Q1 and Q2 with their preferred instrument

The graph below shows the most played instruments in the most connected nodes of the network. We notice that the saxophone is largely in the lead with 12 occurrences, followed by the piano and the trumpet with respectively 9 and 8 occurrences. It also appears that the two least represented instruments in the most connected nodes are drums and bass.

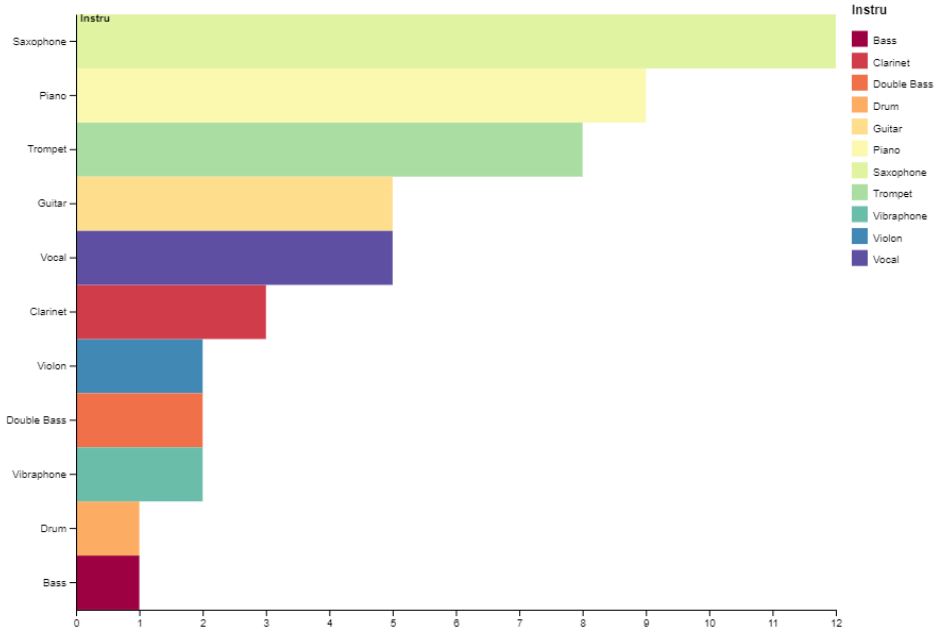

Figure 17: Most played instrument in the top 50 nodes

For the sample of musicians whose degree is between Q1 and Q2 (*Figure 18*), the double bass is the most popular instrument with 7 occurrences. Then come, as before, the trumpet and the piano with both 6 occurrences. The saxophone which was the most popular instrument for the most connected musicians is here in fifth position with 4 occurrences. It also seems that drums and trombone are slightly more popular in this sample than in the previous one.

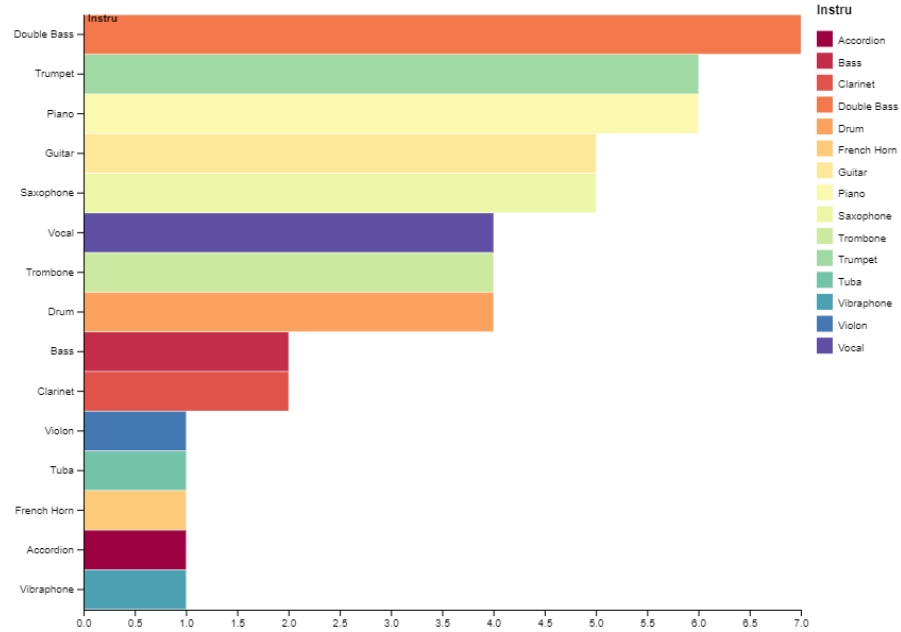

Figure 18: Most played instrument in the 50 nodes randomly selected between Q1 and Q2.

The instrumental compositions of a jazz band depends heavily on the style of jazz played by the band and by the number of musicians. There are several jazz ensemble compositions, the most common are duets, trios, quartets, quintets, sextets and beyond 12 musicians we generally speak of Big Band or orchestra. All these compositions can take very different forms. The most common form of trio in jazz includes a pianist, a double bass player and a drummer (e.g. The Bad Plus). A wind instrument such as a saxophone or a trumpet is often added to this form of band.

Several hypotheses could explain the popularity of the saxophone among the most connected musicians. First, the fact that *Sidney Bechet* popularized it in the 1920s to the point of making it indistinguishable from jazz

may have contributed to it being the most represented instrument in the collaborations. Secondly, it could be the soloist and spectacular side of the saxophone that made it the favorite of the most connected nodes. There are many possible explanations for this result and they deserve to be studied in a possible continuation of this document.

It appears that the double bass, which is the most practiced instrument in the randomly picked nodes between Q1 and Q2, is almost not practiced (2 occurences) in the most connected nodes. These results could lead us to think that the double bass players would have more tendency to play with the same musicians or to do less collaborations contrary to the saxophone players who would have more tendency to collaborate. However, nothing allows us to think that it is the instrument that influences the number of collaboration of a musician; other factors must be taken into account.

Finally, it can be noted that there is a greater diversity of instruments in the sample of less connected nodes. There are 15 different instruments in this group versus 11 in the group of highly connected nodes. Thus, it seems that the hubs of the network are be more likely to play the same instruments in contrast to the less connected nodes.

5.3.2 Geographic location

The graph below shows the number of high degree nodes in relation to their birthplace (city and country). What stands out the most is the predominance of the United States of America in this graph with 82% of the 50 most connected musicians coming from this country. It also appears that New York City has been the birthplace of most collaborative jazz musicians.

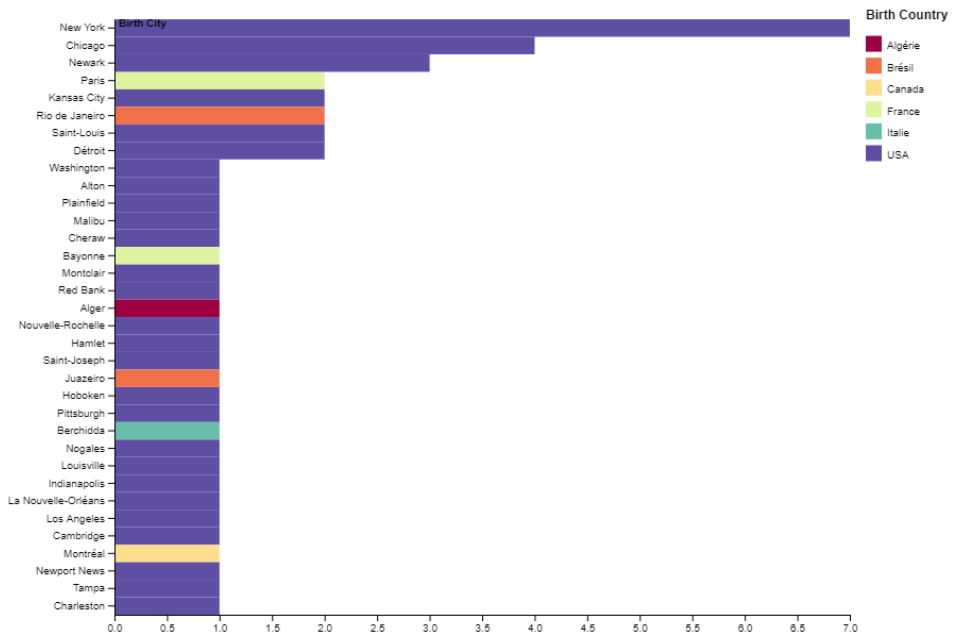

Figure 19: Most common birthplace of the top 50 nodes

Regarding the 50 nodes randomly selected between Q1 and Q2, we notice that the United States of America is still the most represented country in the sample (with 34 occurrences), representing 68% of the total. In addition, countries that were not present in the sample of most connected nodes such as Argentina, United Kingdom, Turkey, New Zealand and South Africa appear in the data. The second most present country in the sample is France with 7 occurrences against 3 in the previous sample.

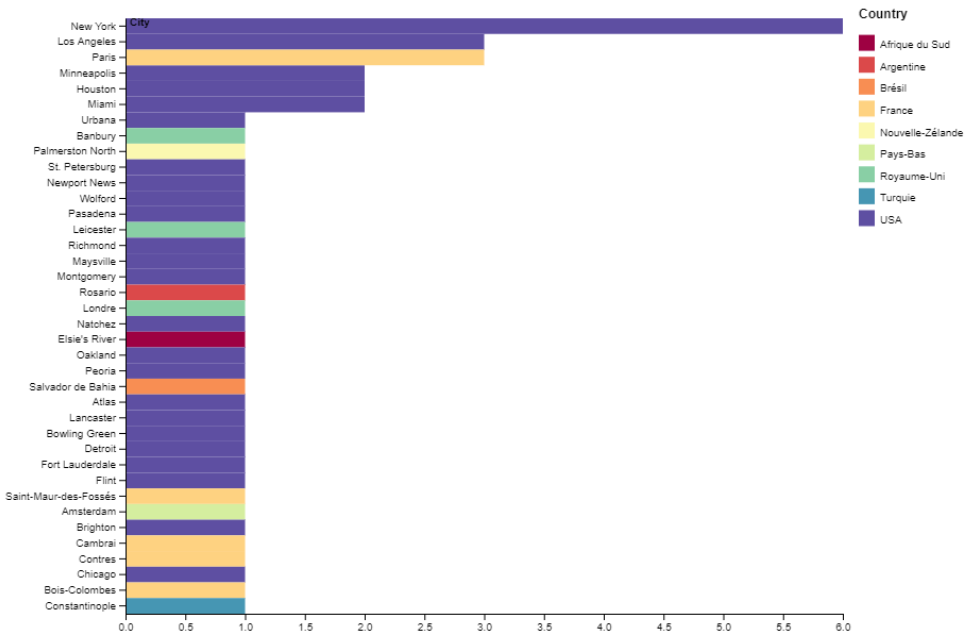

Figure 20: Most common birthplace of the 50 nodes randomly selected between Q1 and Q2.

For the two samples analyzed, it appears that the United States of America and more precisely New York is the birthplace of the greatest number of jazz musicians.

It is generally accepted that the geographical origin of jazz is in the United States of America and more precisely in New Orleans and Louisiana. This origin could explain the significant presence of the United States in the countries of birth of the most connected nodes. Moreover, the predominance of New York as the birthplace of the most connected musicians seems to be explained by the fact that the most famous jazz clubs were located in New York. Among others, the *Savoy Ballroom* and the *Cotton Club* are both in Harlem. Moreover, the strong presence of Paris (France) seems to be explained by the fact that in Europe, this city is one of the first, under the influence of the U.S.A. after the first war, where Jazz is diffused in the clubs and cafés, such as the *Hanneton* or the Blue Note.

5.3.3 Year of birth

It seems important to specify that the age of the musicians and the age of the node in the network are two distinct parameters. For example, a musician might start collaborating (publicly) at an older age, so the node representing him or her will be recent and thus the age of that node will be low. However, the older a musician is, the more likely he or she is to join the network early in its construction and thus have an older node as well. This section will deal with the actual age of the musicians, not the nodes that represent them.

It appears that all the 50 most connected nodes were born between 1897 and 1980, the oldest node (*Sidney Bechet*) would be 124 years old today (2021) while the youngest is 58 (*Dave Douglas*) years old. The current average age for these 50 nodes is 97.28 years. The only two nodes with a degree greater than 400 have an average age of 108.5 years. Furthermore, the 10 most connected nodes, with an average degree of 305.4, have an average age of 95.8 years.

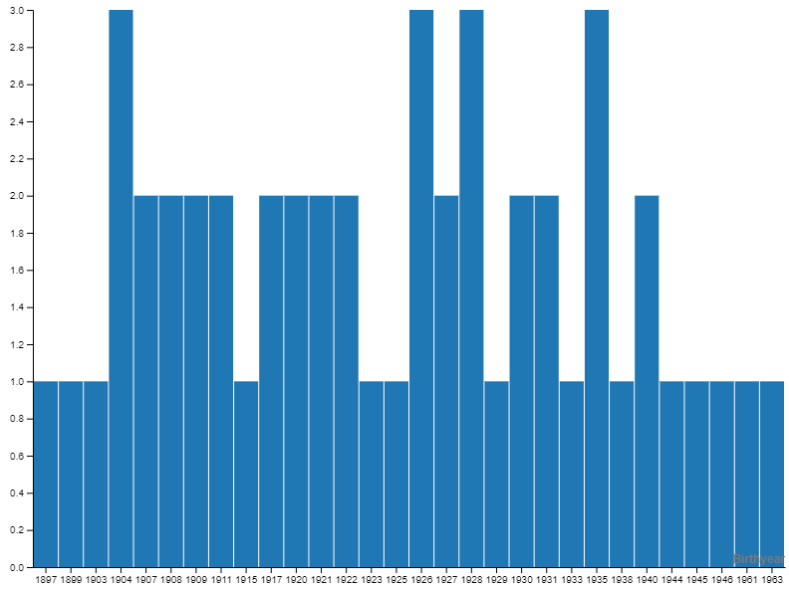

Figure 21: Year of birth of the 50 most connected nodes

In the second sample studied (*Figure 21*), it seems that all nodes are born between 1891 and 1982, the oldest node (*Porter Grainger*) would be 130 years old today (2021) while the youngest (*Johnathan Finlayson*) is 39 years old. The current average age for these 50 nodes is 79.36 years.

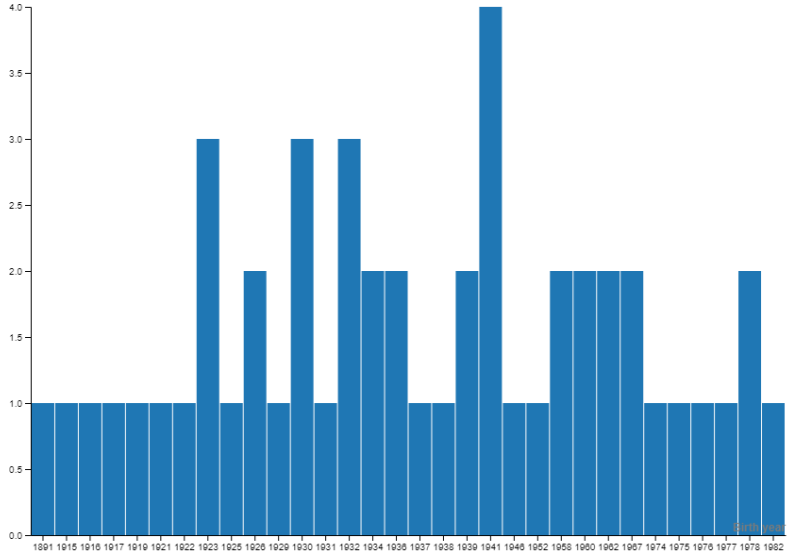

Figure 22: Year of birth of the 50 nodes randomly selected between Q1 and Q2.

We notice that the average age of the most connected nodes is higher than the average age of the second sample, the difference between the two averages is 17.64 years. As empirical results have shown, it seems that age of the node is an important factor in being a highly connected node. As previously indicated, the age of the musician is often linked to the age of the node that represents him. This may be explained by the fact that the older a musician is, the more time he or she has had to play with other musicians (i.e. make early collaborations).

5.3.4 Labels

This section will discuss the labels under which most of the songs of the musicians in the two samples analyzed were released.

Regarding the sample with the most connected musicians, it seems that the label that has produced/accompanied the most of top nodes musicians are *Verve* (17), *Columbia* (16) and *Blue Note* (14). According to the data

collected, there are 62 different labels for this sample. It is important to note that a musician may have worked with several labels during his or her career.

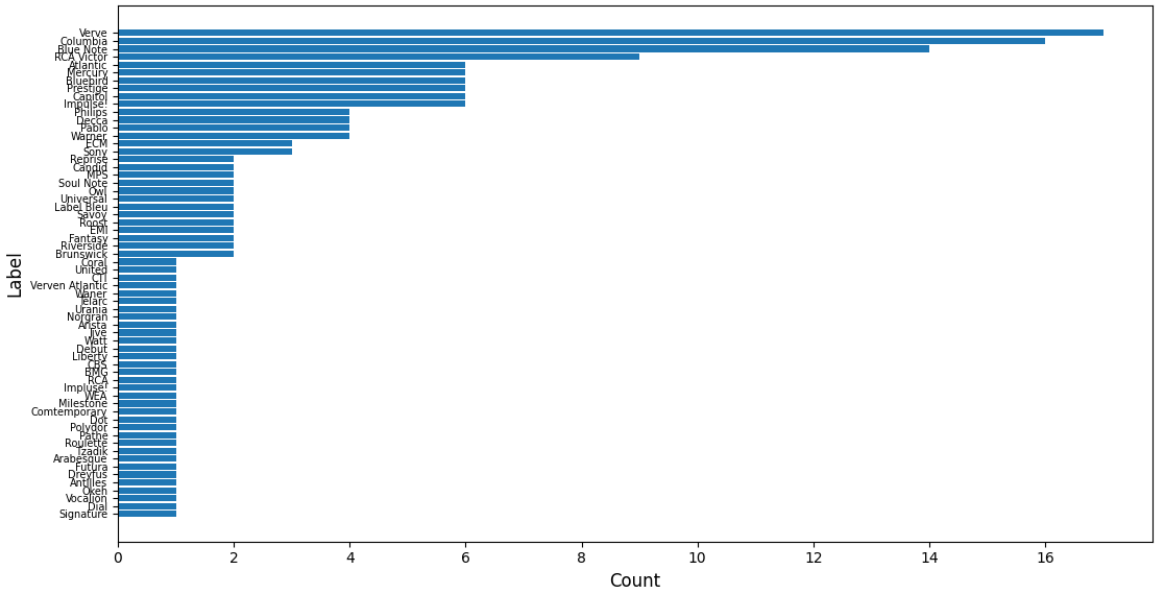

Figure 23: Number of most connected nodes by label

According to the data collected, there are 70 different labels for this sample and it seems that the label that has produced/accompanied the most of musicians between Q1 and Q2 are *Blue Note* (8), *Columbia* (7), *ECM* (6) and *Verve* (6).

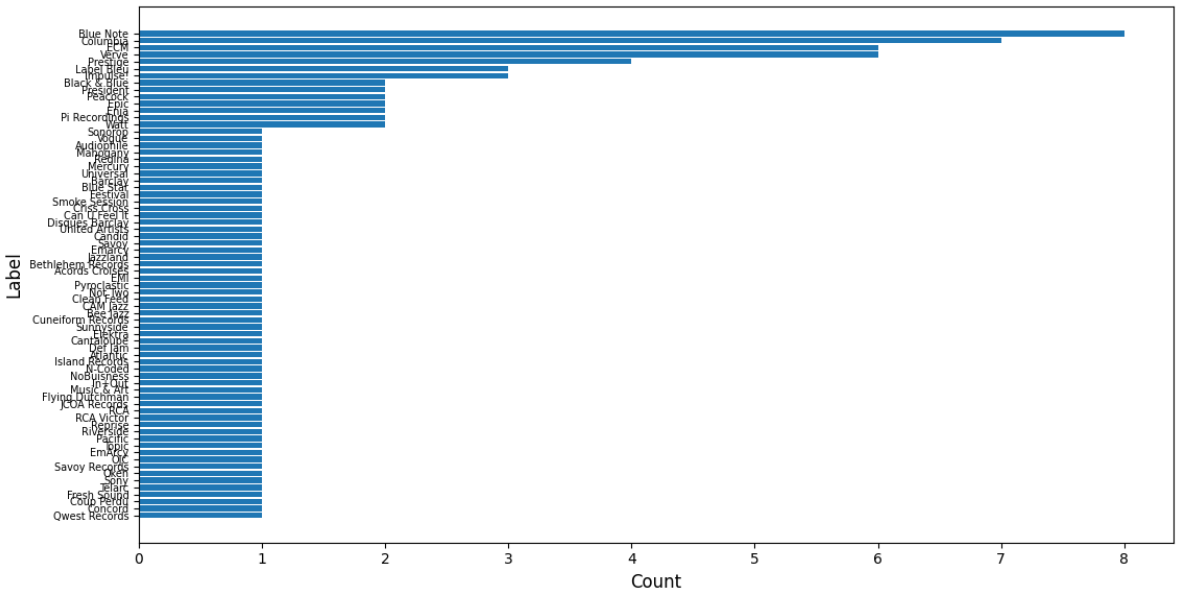

Figure 24: Number of nodes between Q1 and Q2 by label

For both ensembles, we note a large diversity in the labels having produced only one musician. It seems that for both samples, the same labels are best represented in the collected data. These are *Columbia*, *Verve*, *Blue Note*, *Prestige* and *Impulse!*. This could mean that the label under which an album is released, or with which a musician works, has no obvious influence on the ability of a node to be or not to be a network hub.

6 Limitations and future improvements

In the possible continuation of this document, the author would like to considerably increase the size of the network and thus the number of musicians composing it. This in order to be able to draw even more reliable conclusions than those presented in this paper. The size of the network is limited by the data available on the various Wikipedia pages used as sources. Like all research work, the one presented here was limited in time. Many interesting and complex questions remain to be explored, for example: Is there a link between gender, ethnicity and being a network hub? What is the proportion of collaborations between musicians from different countries? Are musicians from the same city more likely to collaborate with each other, as is the case for musicians from the same country?

A parameter that seems important to take into account and that has not been dealt with in this paper is the popularity of musicians. A question that comes to mind is whether the popularity of a musician influences his place in the network (i.e. being a hub or not). This parameter could possibly be quantified by looking at the sales figures for the albums and possibly the number of people present at the different concerts. Another possible future improvement, mentioned in the point 4.3.1, which concerns rather the practical and technical parts of the project, would be to improve the time taken by the program to build all the links between the musicians and to generate the network visually. Another parameter that can be added to the data in the future and that has not been retained here for reasons explained in section 4.2 is the gender of the musicians. It would be interesting to see if there is, for example, an evolution, over the last few years, of the place of women in the world of jazz, like other styles, which is unfortunately known for its lack of gender equity. In a future research, it could also be interesting to look at the different musical sub-genres included in jazz. In these pages, Bebop, Jazz-Fusion, Bossa Nova, etc. have been considered as one and same music, which for jazz specialists is considered as an unforgivable mistake. Thus, the links between the musical sub-genre, the communities, the hubs and the geographical origins, for example, would be explored. One other future improvement that the author would like to bring to this document would be to implement the various methods allowing the detection of community in order to compare the obtained results and to vary the types of methods (e.g. superposition of communities, dynamic, ...). Finally, as explained in section 4.2, a future improvement would consist in automating, with the help of the NLP aglorithm, the recognition of different pseudonyms referring to a same musician.

7 Conclusion

Regarding the topology of the network studied in this paper, it appears that the network is of the dense type, the nodes composing it being well interconnected. It also appears that the most connected nodes (hubs) of the network are strongly connected to each other, which could translate into a certain robustness. Thus, if we remove some hubs, the connectivity of the network would not be significantly affected. According to the algorithm used to detect communities, it seems that they are mainly delimited by a geographical factor. Finally, concerning the topology, it appears that the studied network corresponds to the characteristics expected for a scale-free network: the value of its gamma corresponds to the expected results and it is the same for the degree distribution of the nodes.

As for the research question introduced at the beginning of this work, the conclusions that can be put forward are the following. Regarding the concept of preferential attachment being inseparable from the concept of hub, it would seem that the information carried by the most connected nodes of the network can provide a first answer to the question addressed at the section 1.1. Thus, as a result of the research and experiments conducted so far in this paper, it would appear that the parameters that facilitate preferential attachment in a collaborative jazz network are multiple:

Firstly, the geographical origin seems to be the most important parameter, as it appears that musicians born in the United States of America are more likely to attract collaborations with other musicians, more precisely, the musicians born in New York. The importance of the United States of America can be explained by the origin of the creation of jazz music.

Second, the instrument of the jazz musicians could be a parameter influencing preferential attachment. It seems that a musician who plays the saxophone is potentially an important factor in attracting new collaborations, although the cause-and-effect relationship is difficult to demonstrate in practice.

Next, the year of birth of musicians also seems to play a role in preferential attachment. Indeed, the older a musician is, the more likely he or she will join the network early in its construction and thus have more time to collaborate with other musicians in the network.

Finally, a last parameter that comes into play in preferential attachment, and that seems almost impossible to study, is the human factor. Indeed, no matter how important the factors favoring preferential attachment are, if two musicians do not get along, if they do not have chemistry, there is almost no chance that they will ever collaborate. This observation is a common problem in the study of all so-called social networks, as is the case here in this collaborative network. Furthermore, the conclusions drawn from the hubs are based on statistical results specific to the network studied, depending on the data sets constructed, so these conclusions provide a general picture of the top hubs. It seems obvious that there will be exceptions that do not fit into the conclusions presented above.

8 Bibliography

[1] M. E. J. Newman. Scientific collaboration networks. ii. shortest paths, wieghted networks and centrality. 10.1103/PhysRevE.64.016132, June 28, 2001.

[2] C. Ducruet L. Beauguitte. Scale-free, small-world networks et géorgraphie. https://halshs.archives-ouvertes.fr/halshs-00601211, June 17, 2011.

[3] R. Albert A. Barabasi. Emergence of scaling in random networks. 10.1126/science.286.5439.509, October 15, 1999.

[4] A. Barabasi R. arlbert. Statistical mechanics of complex networks. 10.1103/RevModPhys.74.47, June 6, 2001.

[5] N. Parmar J. Uszkoreit L. Jones A. N. Gomez L. Kaiser I. Polosukhin A.Vaswani, N. Shazeer. Attention is all you need. https://arxiv.org/pdf/1706.03762.pdf, December 6, 2017.

[6] L. Danon P.M. Gleiser. Community structures in jazz. 10.1142/S0219525903001067, October 1, 2003.

[7] M.E.J. Newman M. Girvan. Community structure in social and biological networks. https://doi.org/10.1073/pnas.122653799, April 6, 2002.

[8] M. Barthélemy S. Fortunato. Resolution limit in community detection. https://www.pnas.org/content/104/1/36, d November 6, 2006.

[9] P. L. Krapivsky E. Ben-Naim. Stratification in the preferential attachement network. https://arxiv.org/abs/0905.1920, May, 2009.

[10] M. Udrescu Radu Marculescu A. Topirceanu. Weighted betweenness preferential attachment: A new mechanism explaining social network formation and evolution. https://www.nature.com/articles/s41598-018-29224-w.pdf, July 18, 2018.

[11] A. L. Barabási H. Jeong, Z. Néda. Measuring preferential attachment for evolving networks. https://iopscience.iop.org/article/10.1209/epl/i2003-00166-9/pdf, 1 February, 2008.

[12] A. Pal S. Ruj. Preferential attachment model with degree bound and its application to key predistribution in wsn. https://arxiv.org/pdf/1604.00590.pdf, 3 April, 2016.

[13] G. Fiumara A. Provetti P. De Meo, E. Ferrara. Generalized louvain method for community detection in large networks. https://ieeexplore.ieee.org/document/6121636, August, 2011.

[14] M. Amad R. Djerbi, R. Imache. Communities' detection in social networks: State of the art and perspectives. 10.1109/ISNCC.2018.8531055, June, 2018.