How to read networks and make them legible

The "jazz network" test bed graph

To exemplify our method, we wanted to use a 'standard graph', but most test bed networks were too small for our purposes – for instance, the famous "Karate Club" of Zachary, 1977 contains only 34 nodes. It is easy to observe relational structures in networks of a few dozens or hundreds of nodes, but we wanted to show that VNA can also be applied to networks with several thousands of nodes. Inspiration came from another graph often discussed in the literature: the network of collaborations between jazz musicians produced by Gleiser & Danon (2003). As observed by the McAndrew et al. (2014), "as a music form, jazz is inherently social" and thus particularly propitious to network analysis. Yet, Gleiser & Danon network contains only 1.473 nodes and is limited to the jazz bands that performed between 1912 and 1940 (making it difficult to interpret for the contemporary reader). We thus decided to produce an updated and expanded "jazz network" by drawing on Wikipedia's ontology. Here is the protocol that allowed us to obtain a graph of 6.049 nodes and 85.842 edges:

- We used Wikidata.org to extract
    1. All the 6.796 'instances' of 'human' and the 976 'instances' of 'band' with 'genre = jazz'. We thus obtained a list of individuals and bands that have a page in the English Wikipedia and that are related to jazz (mostly jazz musicians, but also jazz historians and producers). For each of them, we also collected (when available):
        - the 'birth year' (for individual) and 'inception' date (for bands)
        - the 'citizenship' (for individuals) and 'country of origin' (for bands) – when multiple nations were available, we kept only the first one.
        - the 'ethnic group' and 'genre' for individuals.
    2. All the 53 'subgenres' of the genre 'jazz' and all the 396 'record labels' associated with the individuals and bands of the list above.

- We used the Hyphe web crawler (hyphe.medialab.sciences-po.fr; Jacomy et al., 2016; Ooghe-Tabanou et al., 2018) to visit all the pages of the elements above in English Wikipedia and extract the hyperlinks connecting them.

- From the resulting network
    - We removed all the edges that did not have an individual or a band as one of their vertices (for reasons that we will discuss later).
    - We kept only the largest connected component (the largest group of connected nodes and edges), obtaining a network of 6.381 nodes (5396 individuals, 589 jazz band, 346 record labels and 50 subgenres) 85.826 edges.

Positioning nodes

In the introduction we argued that the most important visual variable of VNA is the position of the nodes. Nodes that are more directly or indirectly associated, we wrote, *tend to* find themselves closer in the spatialised network. The caution introduced by "tend to" is crucial, because (as we will show in section 4), there is no strict correlation between the geometric distance in the spatialised graph and the mathematical distance (however defined) in the graph matrix. In VNA, it is not the exact position of any specific node that should be considered, nor the distance between node couples, but the general grouping of nodes and the disposition of such groups. It is not the nodes' position that counts, but the *nodes' density*. In particular, what should catch the eye of the observer are empty spaces.

In a continuum that goes from a set of disconnected nodes to a fully connected clique, the structure of a network is defined by the full and the voids created by the uneven distribution of its relations. Since force-directed layouts would represent both extremes as circles filled with nodes placed at the same distance, everything that departs from this disposition is an indicator of structure. When analysing a spatialised network, therefore, look for shapes that are not circular – which indicate polarisation – and of difference in the density of nodes – which indicates clusterisation.

Don't be too quick discouraged, however, if your network looks like look like amorphous tangle (a 'hairball' as in network jargon). The legibility of network visualisations depends crucially on the choice of the spatialisation algorithm. Though all force-directed algorithms are based on a similar system of attraction and repulsion forces, their results may differ because of the specific way in which they handle computational challenges (in particular optimisations necessary to reduce calculations) and visual problems (in particular the balance between the compactness and legibility). What can, at first, be mistaken for a homogenous distribution of connections can, in some case, derive from an unfortunate choice of the spatialisation algorithm or its settings.

This is why, among the many tools available for network analysis, we recommend Gephi (gephi.org, Bastian et al., 2009) and Sigma.js (sigmajs.org). Having been developed expressly for networks drawing, these pieces of software do *not* treat spatialisation as an automated operation but offer a subtle control of visual variables. Among the force-directed algorithms our favourite is ForceAtlas2, because it offers good performances on relatively large networks while implementing attraction and repulsion in a relatively pure way (cf. Jacomy et al., 2014).

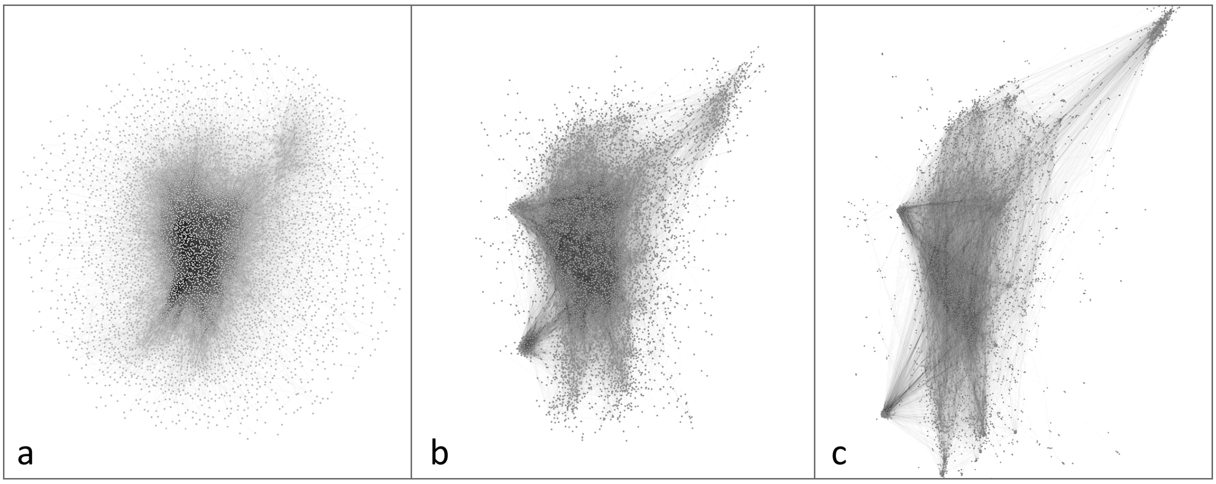

Figure 1. The 'jazz network' spatialised (a) with the algorithm proposed by Fruchterman & Reingold, 1991, (b) with ForceAtlas2 (with default parameters) and (c) with ForceAtlas2 with tweaked parameters for 'LinLog mode' and 'gravity'

As an example, the image above shows how our network of jazz individuals and bands (for the moment, we are filtering out subgenres and record labels) look as a hairball when spatialised with Fruchterman and Reingold algorithm (considered as the first computer implementation of force-directed layout, see Fruchterman & Reingold, 1991), but acquire a clearer structure when visualised with ForceAtlas2, particularly when two crucial parameters are adjusted.

The 'LinLog mode' parameters tweaks the way in which distance is taken into consideration in the computation of attraction and repulsion forces. In default ForceAtlas both forces are linearly proportional to the distance (with inverse for attraction), but, as demonstrated by Noack (2009), using a logarithmic proportionality for repulsion makes clusters more visible. 'Gravity', on the other hand, is a generic force that pulls all nodes toward the centre. While it avoids disconnected nodes to drift infinitely far from the rest of the network, such a gravitational force interferes with the purity of force-directed layouts (if too high gravity packs all the nodes in the centre of the space). Activating the LinLog mode and setting the gravity to zero tends to make the clusters more visible, but also produce a more scattered network. As a consequence, it is impossible to suggest a 'catch-all' setting for these parameters. Recursively adjusting the spatialisation parameters to the analysed networks is crucial to make the relational structures visible (just as choosing the right chart and tweaking its visual properties is essential to make sense of a large data table).

Sizing nodes and labels

Now that we have positioned the nodes of our network, in order to reveal effects of polarisation and clustering, we still have to make sense of what we see. To do so, VNA draws on two ancillary visual variables (Bertin, 1967): size and colour. Let's consider size first.

Tools like Gephi allow to change diameter of the points representing the nodes according a variable selected by the user. 'Degree' (the number of edges connected to a node) or, in directed networks, the 'in-degree' (the number of *incoming* edges) are classic choices, as they represent a classic translation of visibility in networks. Being entirely relational, degree can be computed for any networks (and any directed networks in the case of in-degree). Yet, when available, other non-relational variables could be equally interesting. For instance, we can change the size of the elements of our networks according to the number of visits that each of the related Wikipedia page received in 2017.

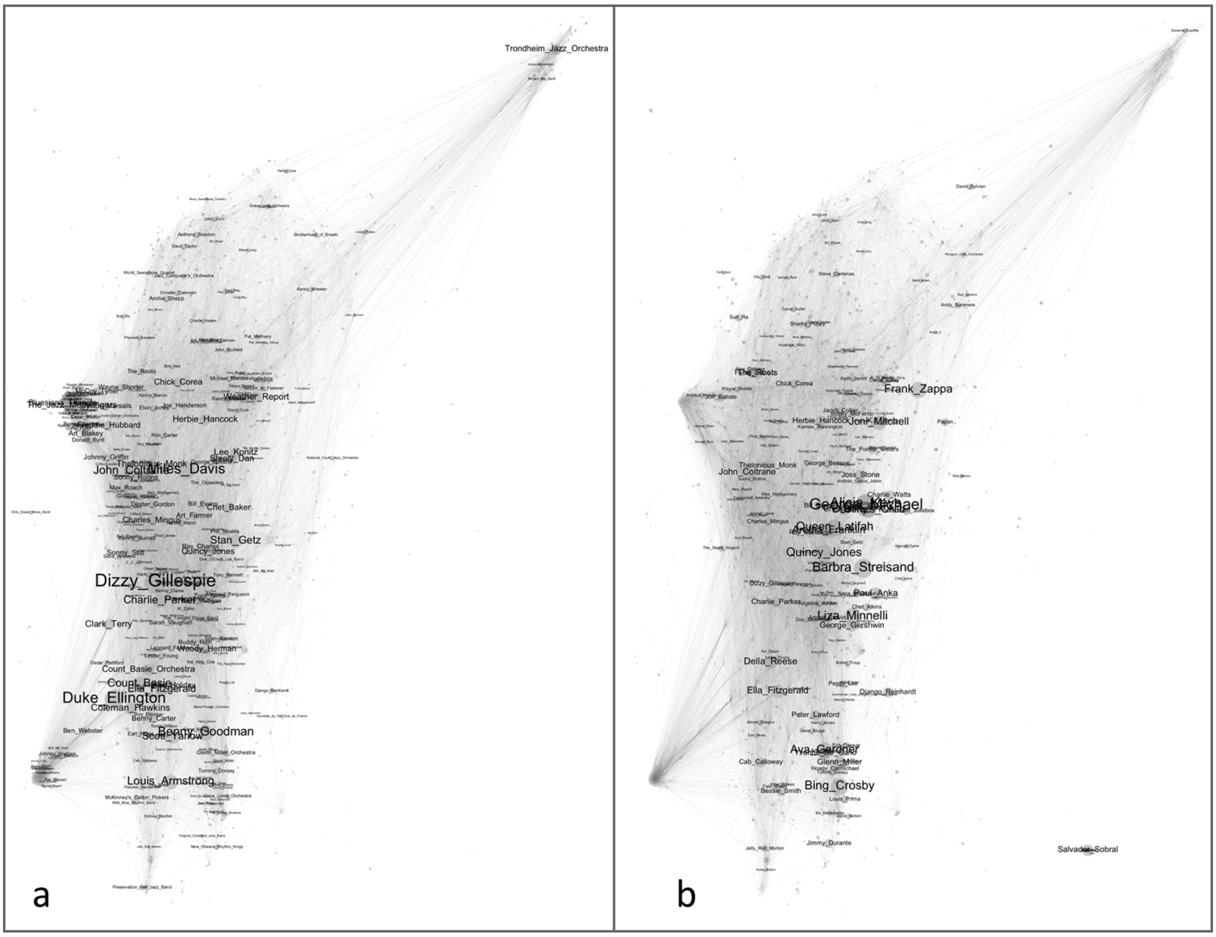

Figure 2. The 'jazz network' with nodes and labels sized according to (a) the in-degree of the nodes of the graph; (b) the number of page views of the related pages in the English Wikipedia.

Note that in the figure 3, we have varied not only the size of the nodes, but also of their label (and even deleted all the labels smaller than a given threshold). This foregrounding operated through size is crucial in VNA because when working with networks with hundreds or thousands of nodes, inspecting all of them is clearly not an option. Changing label size (and dropping some labels), however, entails losing some information, and this is why using more than one scaling variable is always advisable.

Observing the labels of the most visible nodes, we can start to make sense of the factors that shape our network. Comparing the two images in figure 3, for example, it is possible to remark that the pages with high in-degree tend to be positioned on the left, while pages with high pageviews are rather found on the right. Also, nodes with high in-degree are all famous jazzmen (the top five being Dizzy Gillespie, Duke Ellington, Miles Davis, Benny Goodman and John Coltrane), while nodes with high pageviews seems to be pop-culture celebrities (the top five being George Michael, Alicia Keys, Barbra Streisand, Liza Minelli, Bing Crosby). This suggests that a left-right polarization may exist corresponding to a difference between a purer jazz lineage and the contamination with other genres.

This polarisation, however, is a weak one, not only between the left and right of the image, but also and most importantly because the network appears to be stretched vertically much more than horizontally. To what may this vertical polarisation correspond?

Colouring nodes

To investigate the vertical polarisation of our jazz network, we will add to position and size a third visual variable – colour. According to Jacques Bertin (1967), colour can be decomposed in two different variables: brightness (or value) which is better suited to represent continuous numerical variables and hue which is better suited to represent categorial variables. VNA makes use of both.

Noticing at the bottom names such as Lois Armstrong, Duke Ellington and Bing Crosby and at the top Chick Corea, Weather Report and Frank Zappa, we can hypothesise that the vertical polarisation of our network is connected to time and in particular to the period in which the different actors were most active in the jazz scene. While such information is not available in our network, we do have the year of birth and of inception of individuals and bands and we can project them on the network using a scale of brightness going from black (for the oldest actors) to white (for the newest).

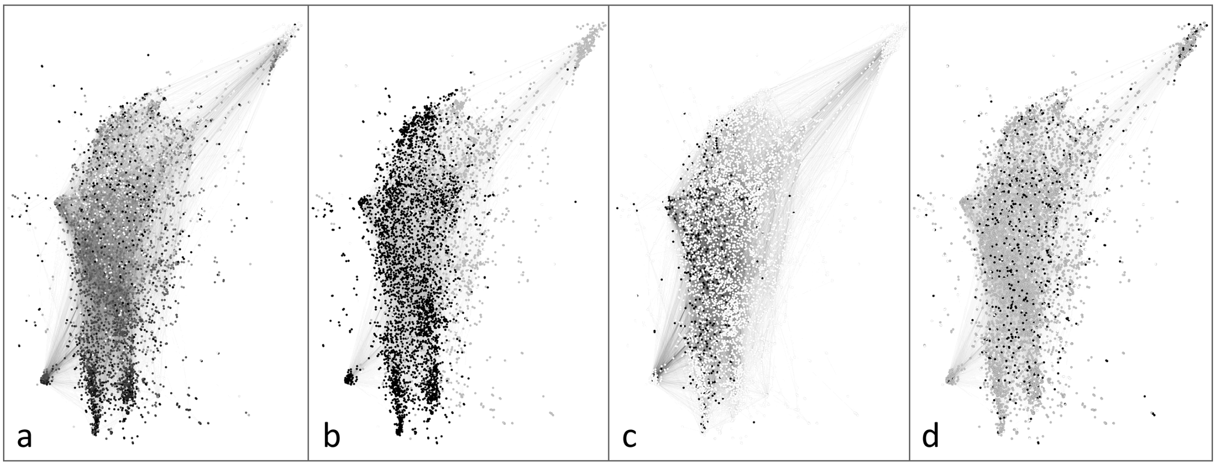

Figure 3. The 'jazz network' with nodes coloured according to (a) the year of their birth or inception (from dark for older individuals and bands to white for newer); (b) their nationality (black for US, grey for all other countries, white for not available); (c) their ethnic groups (black for African American, grey for other ethnic groups, white for not available); (d) their genre (black for women, grey for men, white for not available or others)

The first image the figure 4 seems to confirm our hypothesis that the vertical polarisation corresponds to time. While the separation is not complete, darker nodes are more present at the bottom of the image and brighter at the top.

In the other three images in figure 4, we relied on hue (using only black, grey and white and no intermediary shades) to observe how different categories distributes in the network. Figure 4b and 4c are dedicated respectively to the nationality and ethnic group. While they are difficult to interpret alone, together they suggest an interpretation. Figure 4b, reveals unsurprisingly that jazz is primarily an American genre of music (but remember that we relied on English Wikipedia to build the network), but it also shows that most non-American actors (in grey) tend to be on the right of the image. Similarly, figure 4c shows that while most nodes are not qualified, the only ethnic group that stands out is African American (again not surprisingly knowing the history of the genre). The nodes representing African American actors (in black) are everywhere in the network, but slightly more to its left than to its right. Both observations seem to confirm the interpretation we got from figure 3, that the horizontal polarisation is loosely connected to the 'purity of the attachment to the jazz genre'.

To be sure, not all variables will turn out to be connected to the visual structures of the network. In figure 4d, for example, we show how genres are completely mixed in our network, in a way that suggests that at least in this field genre does not produce a relational fracture (but notice how men are significantly more numerous than women).

Using force-directed spatialisation to determine the position of nodes and size and colour to project various variable on our visualisation, we have identified two perpendicular axes of polarisation of our jazz network (with a main vertical axis defined by time and a secondary horizontal axis defined by 'genre purity'). This configuration is distinctive of this network and is not to be expected in every network. Other networks can have a single axis of polarisation, more than two and sometimes none (being instead are 'stretched' between multiple poles).

Naming clusters

So far, we have looked only at the poles of our graph, not at its clusters. We have considered the shape of the network, but not the different zones of density produced by the disposition of nodes. In VNA clusters are defined as regions that gather by many nodes closely packed together and surrounded by areas with a much sparser density (the "structural holes" of Burt, 1995).

In the jazz network, the only easily identifiable cluster is the one located at the very top right of the image and whose most visible node is the Trondheim Jazz Orchestra (see figure 3), which contains a group of mostly Norwegian musicians most of which are members of the Orchestra. The other clusters of our network are more difficult to identify and make sense of. To do so, we present in this paper two advanced techniques for visual network analysis. These techniques facilitate, but do not replace the basic operation of thoroughly examining the density and reading nodes labels and qualification (when available) to make sense of why some groups of nodes are more closely connected than others.

The first technique entails is not available in Gephi but can be performed through another tool called Graph Recipes (tools.medialab.sciences-po.fr/graph-recipes) and based on Sigma.js. Using a special script available (as all the scripts that we used to create the network and the networks itself) at www.tommasoventurini.it, we transformed our network in an heatmap in order to make the differences of density more salient (see figure 5).

The second technique entails qualifying the different areas of the network using 'qualifying nodes'. This technique consists in adding to the network a new set of nodes that do not influence the spatialisation but can be used to make sense of it. In our example, we used the subgenres of the genre jazz (according to Wikidata) and the record labels associated with the artists and ensembles of our network. To make sure that these qualifying nodes do not influence the layout, we used a 'double spatialisation'. We first spatialised the network with the only (of the individuals and the bands). We then froze the position of these 'primary nodes', added the subgenres and record labels and run the spatialisation algorithm a second time on the qualifying nodes only. A last detail: though the Wikipedia pages related to the subgenres and record labels have hyperlinks connecting them, we have removed these edges from our network, so that the qualifying nodes are only position according to their connections to the primary nodes (and not according to the connections between themselves).

After the double spatialisation, the qualifying nodes can be used to suggest labels for the clusters of the networks in which they end up being located. To complete our visualisation, we worked with a jazz expert (Emiliano Neri, whom we heartfully thank for his help), to drop most primary and qualifying labels and keep only the most significant.

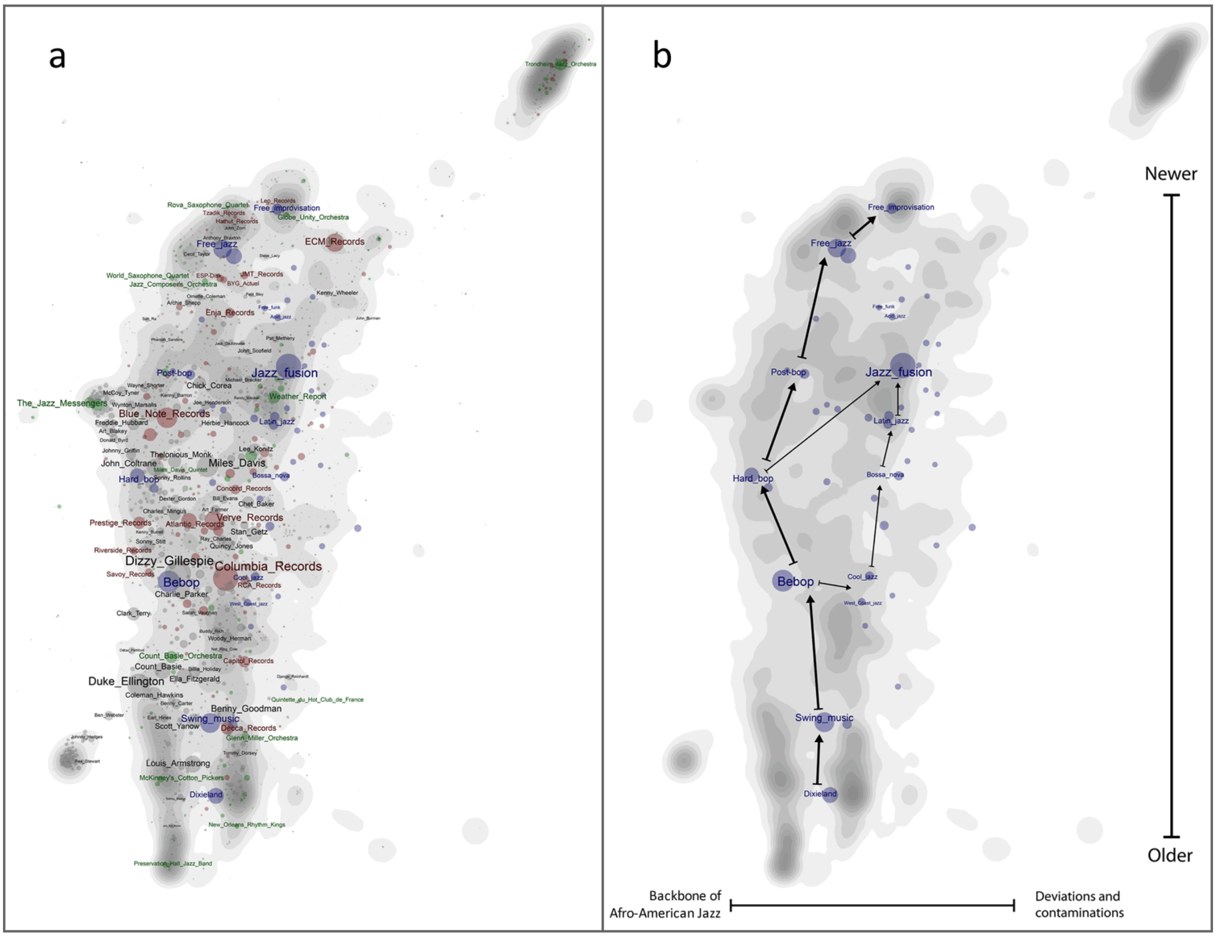

Figure 4. The 'jazz network' with (a) the labels of the most salient node of each type (grey for individual, green for bands, blue for subgenres and red for record labels) and (b) the identification on the structure of the network in terms of the evolution of the jazz musical language.

Interpreting the position of nodes and clusters

Now that we have decided on how to spatialize the network, how to size and colour it's nodes, and how to name its clusters, we can try to make sense of both its overall structures and of the position of its most important nodes. As we will argue in the next section, it is a distinctive advantage of VNA that it allows observing global patterns and local configurations in the same visual space.

In figures 6 and 7, one can observe (moving from the bottom to the top of the image) the development of jazz musical language. This evolution occupies the left of the image and starts from *dixieland* and *swing music* and progresses to *bebop*, *hard bop*, *post bop* and finally to *free jazz* and *free improvisation*. From this backbone of Afro-American jazz, depart on the right of the charts deviations (such as the *cool jazz* and *west coast jazz*) and contaminations with other genres (such as *bossa nova* , *latin jazz* and later *jazz fusion*).

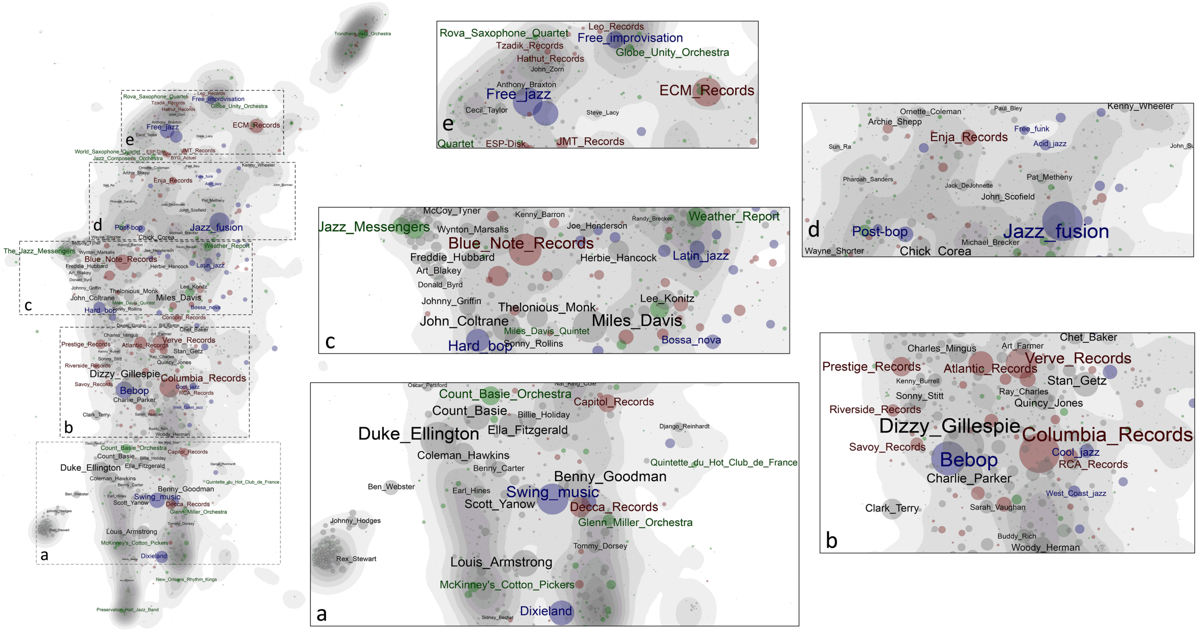

Figure 5. Mosaic providing a zoom on the different regions of the 'jazz network'

[7.a] The bottom of the image corresponds thus to the early years of the genre and is marked by Decca Records, a label which dominated the jazz scene in the 1930s and 1940s, and Capitol Records, also particularly active in the 1940s. The region of *dixieland* and *swing music* is split in two parallel clusters (already identified by Glaiser et al., 2003): to the right, the and 'white big bands' gathered around *Tommy Dorsey*, *Glenn Miller* and *Benny Goodman*; and to the left the 'black big bands' gathered around *Louis Armstrong*, *Coleman_Hawkins*, *Count Basie* and, *Duke Ellington*. This last bandleader is also at the origin of the smaller cluster to the bottom left, constituted by the members of its orchestra. Famous vocalists such as *Ella Fitzgerald* and *Billy Holiday* are positioned toward the centre because of the large number of their collaborations. More to the right, is *Django Reinhardt*, the Romani guitarist, whose isolated position is justified by his living in in Europe.

[7.b] Shifting up toward the *bebop*, many new record labels emerge such as *Prestige*, *Riverside*, *Savoy*, *Atlantic*, and more importantly *Verve* and *Columbia* which were destined to impose themselves in the jazz scenes for years to come. Very close to the node representing *bebop*, one can find (not surprisingly) the trumpeter *Dizzy Gillespie* and the saxophonist *Charlie Parker*, who were among the most influential artist of this new style, and the vocalist *Sarah Vaughan* who collaborated with both. In a more bridging position are *Woody Herman* and *Clark Terry*, whose long careers spanned between *swing* and *bebop*.

[7.c] Move upward, the increase in the number and dispersion of nodes illustrates the growing diversification in jazz language in the 1950s. On the one hand (on the left of chart), *bebop* evolves into *hard bop*, thanks to the *Blue Note* record label and to musicians such as *Charles Mingus*, *Sonny Rollins*, *Thelonious Monk* and *Art Blakey*. This last bandleader is at the origin of the important ensemble of the *Jazz Messengers*, which creates a little cape on the left of the map and which acted as an incubator for talent, including *Freddie Hubbard*, *McCoy Tyner* and *Wynton Marsalis*. On the other hand (on the right of the chart), the experiences of west coast jazz and cool jazz evolve through the contamination with styles from Latin America, giving birth to *bossa nova* and *latin jazz*, popularized in the US by influential figures such as *Stan Getz* and *Quincy Jones. John Coltrane* and *Miles Davis* occupy the centre of this region (and of the whole graph) for the crucial role they played in bridging all these experiences.

[7.d] In the 1960s, the contaminations observed in the centre-right of the chart turn toward rock and funk music as well as their use of electric instruments and amplifiers, originating the so-called *jazz fusion*. Musicians such as *Chick Corea*, *Herbie Hancock*, *John Scofield* and *Pat Metheny*, as well as the group *Weather Report*, play a crucial role in this experience. At about the same time, and with connections assured by artists such as *Joe Henderson* and *Michael Brecker*, *hard bop* develops into *post-bop* thanks to musicians such as *Wayne Shorter* and *Elvin Jones*.

[7.e] In the 1970s, experiences of radical improvisation developed in the previous decades conquered the musical avant-garde, giving birth to *free jazz* and *free improvisation*. Initiated by musicians such as *Sun Ra*, *Cecil Taylor*, *Archie Shepp* and *Ornette Coleman*, this style has been developed by *Anthony Braxton*, *John Zorn*, *Evan Parker* and many others. Interestingly, this genre seems to be edited particularly by European record labels such as *JMT* and *ECM*. This last record label is also the bridge that connects the relatively marginal cluster of the Scandinavian jazz (at the top-right of the figure) to the rest of the maps.