# Node Attribute Analysis for Cultural Data Analytics: a Case Study on Italian XX-XXI Century Music

Michele Coscia[1*]

[1*]CS Department, IT University of Copenhagen, Copenhagen, Denmark.

Corresponding author(s). E-mail(s): mcos@itu.dk;

**Abstract**

Cultural data analytics aims to use analytic methods to explore cultural expressions – for instance art, literature, dance, music. The common thing between cultural expressions is that they have multiple qualitatively different facets that interact with each other in non trivial and non learnable ways. To support this observation, we use the Italian music record industry from 1902 to 2024 as a case study. In this scenario, a possible research objective could be to discuss the relationships between different music genres as they are performed by different bands. Estimating genre similarity by counting the number of records each band published performing a given genre is not enough, because it assumes bands operate independently from each other. In reality, bands share members and have complex relationships. These relationships cannot be automatically learned, both because we miss the data behind their creation, but also because they are established in a serendipitous way between artists, without following consistent patterns. However, we can be map them in a complex network. We can then use the counts of band records with a given genre as a node attribute in a band network. In this paper we show how recently developed techniques for node attribute analysis are a natural choice to analyze such attributes. Alternative network analysis techniques focus on analyzing nodes, rather than node attributes, ending up either being inapplicable in this scenario, or requiring the creation of more complex n-partite high order structures that can result less intuitive. By using node attribute analysis techniques, we show that we are able to describe which music genres concentrate or spread out in this network, which time periods show a balance of exploration-versus-exploitation, which Italian regions correlate more with which music genres, and a new approach to classify clusters of coherent music genres or eras of activity by the distance on this network between genres or years.

**Keywords:** cultural data analytics, complex networks, data clustering, temporal analysis

# 1 Introduction

Node attribute analysis has recently been enlarged by the introduction of techniques to calculate the variance of a node attribute [1], estimate distances between two node attributes [2], calculating their Pearson correlations [3], and cluster them [4] without assuming they live in a simple Euclidean space – or learnable deformation thereof.

These techniques are useful only insofar the network being analyzed has rich node attribute data, and that analyzing their relationships is interesting. This is normally the case in cultural analytics, the use of analytic methods for the exploration of contemporary and historical cultures [5, 6]. Example range from archaeology – where related artifacts have a number of physical characteristics and can be from different places/ages [7–9] –; to art history – where related visual artifacts can be described by a number of meaningful visual characteristics [10–12]; to sociology – where different ideas and opinions distribute over a social network as node attributes [13, 14] –; to linguistics – with different people in a social network producing content in different languages [15]; to music – with complex relations between players and informing meta-relationships between the genres they play [16, 17].

In this paper we aim at showing the usefulness of node attribute analysis in cultural analytics. We focus on the Italian record music industry since its beginnings in the early XX century until the present time. We build a temporally-evolving bipartite network connecting players with the bands they play in. For each band we know how many records of a given genre they publish, whether they published a record in a given year, and from which Italian region they originate – all node attributes of the band. By applying node attribute analysis, we can address a number of interesting questions. For instance:

1. How related is a particular music genre to a period? Or to a specific Italian region?
2. Is the production of a specific genre concentrated in a restricted group of bands or generally spread through the network?
3. Does clustering genres according to their distribution on the collaboration network conform to our expectation of meta-genres or can we discover a new network-based classification?
4. Can we use the productivity of related bands across the years as the basis to find eras in music production?

The music scene has been the subject of extensive analysis using networks. Some works focus on music production as an import-export network between countries [18]. Other model composers and performers as nodes connected by collaboration or friendship links [16, 19–22]. Studies investigate how music consumption can inform us about genres [17] and listeners influencing each other [23–25]. Differently from these studies, we do not focus on asking questions about the network structure itself. For our work, the network structure is interesting only insofar it is the mediator of the relationships

between node attributes – the genres, years, and regions the bands are active on –, rather than being the focus of the analysis.

This is an important qualitative distinction, because if one wanted to perform our genre-regional analysis on the music collaboration network without our node attribute analysis, they would have to deal with complex n-partite objects – a player-band-year-genre-region network – which can become unwieldy and unintuitive. On the other hand, with our approach one can work with a unipartite projection of the player-band relationships, and use years, genres, and regions as node attributes, maintaining a highly intuitive representation.

Deep learning techniques and specifically deep neural networks can handle the richness of our data [26–29]. These approaches can attempt to learn, e.g., the true non-Euclidean distances between genres played by bands [30, 31]. The problem is that this learning is severely limited if the space is defined by a complex network [32], as is the case here. Therefore, one would have to use Graph Neural Networks (GNN) [33–35]. However, GNNs focus on node analysis [36–39], usually via finding the best way of creating node embeddings [40, 41]. GNNs only use node attributes for the purpose of aiding the analysis of nodes rather than analyzing the attributes themselves [42–47]. Previous research shows that, when focusing on node attributes rather than on nodes, the techniques we use here are more suitable than adapting GNNs developed with a different focus [4].

Another class of alternative to deal with this data richness is to use hypergraphs [48] and high order networks [49–52]. With these techniques, it is possible to analyze relationships involving multiple actors at the same time – rather than only dyadic relationships like in simpler network representations – and encode path dependencies – e.g. using high order random walks where a larger portion of the network is taken into account to decide which node to visit next [53, 54]. While a comparative analysis between these techniques and the ones used in this paper is interesting, in this paper we exclusively focus on the usefulness of techniques based on node attribute analysis. We leave the comparison with hypergraphs and high order networks as a future work.

Our analysis shows that the node attribute techniques can help addressing a number of interesting research tasks in cultural data analytics. We show that we are able to describe the eclecticism required by music genres – or expressed in time periods –, by how dispersed they are on the music network. We can determine the geographical connection of specific genres, by estimating their correlation not merely based on how many bands from a specific region play a genre, but how bands not playing that genre relate with those that do. We can create new genre categories by looking at how close they are to each other on the music network. We can apply the same logic to discover eras in Italian music production, clustering years into coherent periods.

Finally, we show that our node attribute analysis rest on some assumptions that are likely to be true in our network – that bands tend to share artists if they play similar genres, in similar time periods, and hailing from similar regions.

We release our data as a public good freely accessible by anyone [55], along with all the code necessary to reproduce our analysis[1].

[1] http://www.michelecoscia.com/?page_id=2336

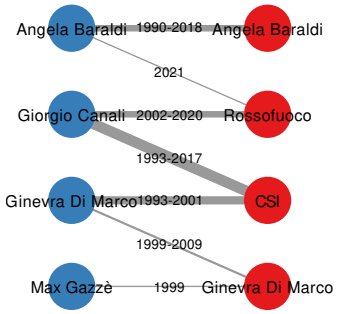



**Fig. 1**: Our bipartite network data model. Artists in blue, bands in red. Edges are labeled with the first-last year in which the collaboration was active. The edge width is proportional to the weight, which is the number of years in which the artist participated to records released by the band.

# 2 Data

In this section we present our data model and a summary description of the data's main features. Supplementary Material Section 1 provides all the details necessary to understand our choices when it comes to data collection, cleaning, and pre-processing.

## 2.1 Data Model

To obtain a coherent network and to limit the scope of our data collection, we focus exclusively on the record credits from published Italian bands. The data from this project comes from crowd-sourced user-generated data. We mainly use Wikipedia[2] and Discogs[3]. We should note that these sources have a bias favoring English-speaking productions. While this bias does not affect our data collection too much, since we focus on Italy without comparing it to a different country/culture, it makes it more likely that there are Italian records without credits, or that are simply missing.

Figure 1 shows our data model, which is a bipartite network $G = (V_1, V_2, E)$. The nodes in the first class $V_1$ are artists. An artist is a disambiguated physical real person. The nodes in the second class $V_2$ are bands, which are identified by their name. Note that we consider solo artists as bands, and they are logically different from the artist with the same name. Note how in Figure 1 we have two nodes labeled "Ginevra Di Marco", one in red for the band and the other in blue for the artist.

Each edge $(v_1, v_2, t)$ – with $v_1 \in V_1$ and $v_2 \in V_2$ – connects an artist if they participated in a record of the band. The bipartite network is temporal. Each edge has a single attribute $t$ reporting the year in which this performance happened. This implies that there are multiple edges between the same artist and the same band, one per year in which the connection existed – for notation convenience, we can use $w_{v_1,v_2}$ to denote this count for an arbitrary node pair $(v_1, v_2)$, since it is equivalent to the edge's weight.

[2] https://it.wikipedia.org
[3] https://www.discogs.com/

| Band | Rock | Pop | Electro | HipHop | Tuscany | Liguria | Piedmont | Veneto | 1998 | 1999 | 2000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Litfiba | 0.448 | 0.038 | 0.008 | 0.000 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| Sabrina Salerno | 0.006 | 0.103 | 0.322 | 0.000 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| Gigi D'Agostino | 0.000 | 0.018 | 0.316 | 0.000 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| Madame | 0.000 | 0.176 | 0.098 | 0.333 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

**Table 1**: A sample of the node attributes. For four sampled bands we show in the tables sections (left to right): the row-normalized number of records released tagged as Rock, Pop, Electronic, or Hip Hop; the one-hot encoded attribute values for the region attribute; the one-hot encoded attribute values for the years of activity.

We have multiple attributes on the band. The attributes are divided in three classes. First, we have genres. We recover from Discogs 477 different genres/styles that have been used by at least one band in the network. Each of these genres is an attribute of the band, and the value of the attribute is the number of records the band has released with that genre. We use $S$ to indicate the set of all genres, and show an example of these attributes in Table 1 (first section). The second attribute class is the one-hot encoded geographical region of origin, with each region being a binary vector equal to one if the band originates from the region, zero otherwise. We use $R$ to indicate the set of regions. Table 1 (second section) shows a sample of the values of these attributes. The final attribute class is the activity status of a band in a given year – with $Y$ being the set of years. Similarly to the geographical region, this is a one-hot encoded binary attribute. Table 1 (third section) shows a sample of the values of these attributes.

## 2.2 Summary Description

For the remainder of the paper, we limit the scope of the analysis to a projection of our bipartite network. We focus on the band projection of the network, connecting bands if they share artists. We do so to keep the scope contained and show that even by looking at a limited perspective on the data, node attribute analysis can be versatile and open many possibilities. Supplementary Section 2 contains summary statistics about the bipartite network and the other projection – connecting artists with common bands.

There are many ways to perform this projection [56–58], which result in different edge weights. Here we weight edges by counting the number of years a shared artist has played for either band. Supplementary Material Section 1 contains more details about this weighting scheme. Since we care about the statistical significance – assuming a certain amount of noise in user-generated data – we deploy a network backboning technique to ensure we are not analyzing random fluctuations [59].

Table 2 shows that the band projection has a low average degree and density, with high clustering coefficient and modularity – which indicate that one can find meaningful communities in the band projection. These are are typical characteristics of a wide variety of complex networks that can be found in the literature.

Table 3 summarizes the top 10 bands according to three standard centrality measures: degree, closeness, and betweenness centrality. Degree is biased by the density of the hip hop cluster – which, as we will see, is a large quasi-clique, including only hip hop bands. Closeness is mostly dominated by alternative rock bands, as they happen to be in the center of mass of the network. The top bands according to betweenness are

| Variable | Band |
|---|---|
| # Nodes | 2,447 |
| # Edges | 6,512 |
| Avg Deg | 5.3 |
| Density | 0.0022 |
| Clustering | 0.4160 |
| Modularity | 0.8437 |

**Table 2**: Summary statistics for the projected networks.

| # | Degree | Closeness | Betweenness |
|---|---|---|---|
| 1 | Night Skinny | Calibro 35 | **Roberto Gatto** |
| 2 | Marracash | Giorgio Canali & Rossofuoco | Adriano Celentano |
| 3 | DJ Double S | Le Luci della Centrale Elettrica | **Vinicio Capossela** |
| 4 | Bassi Maestro | **Vinicio Capossela** | **Luca Carboni** |
| 5 | Jake La Furia | Dente | Marcello Giombini |
| 6 | Noyz Narcos | Afterhours | Pietro Umiliani |
| 7 | Fabri Fibra | **Roberto Gatto** | Renato Sellani |
| 8 | Emis Killa | Elisa | Cube |
| 9 | Gemitaiz | **Luca Carboni** | Tullio Pane |
| 10 | Salmo | Gianni Maroccolo | Ennio Morricone |

**Table 3**: The top 10 bands in the band projection according to different centrality measures. In bold we have nodes central in multiple measures.

those bands that are truly the bridges connecting different times, genres, and Italian regions. Note that we analyze the network as a cumulative structure, therefore these centrality rankings are prone to overemphasize bands that are in the central period of the network, as they naturally bridge the whole final structure. In other words, it is harder to be central for very recent or very old bands.

We visualize the band projection to show visually the driving forces behind the edge creation process: temporal and genre assortativity. For this reason we produce two visualizations. First, we take on the temporal component in Figure 2. The network has a clear temporal dimension, which we decide to place on a left-to-right axis in the visualization, going from older to more recent.

Second, we show the genre component in Figure 3, which instead causes clustering – the tendency of bands playing the same genre to connect to each other more than with any other band. For simplicity, we focus on the big three genres – pop, rock, and electronic – plus hip hop, since the latter creates the strongest and most evident cluster notwithstanding being less popular than the other three genres. For each node, if the band published more than a given threshold records in one of those four genres, we color the node with the most popular genre among them. If none of those genres meets the threshold, we count the band as playing an "other" generic category.

This node categorization achieves a modularity score of 0.524, which is remarkably high considering that it uses no network information at all – and it is not a given that this is the correct number of communities. This is a sign that the network is

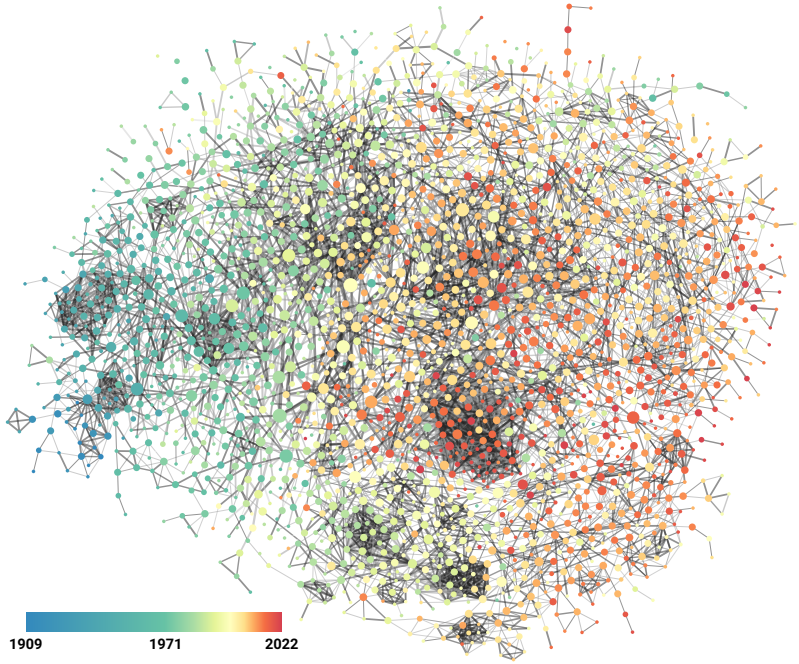



**Fig. 2**: The temporal component of the band projection. Each node is a band. Edges connect bands with significant number of artist overlap. The edge's color encodes its statistical significance (in increasing significance from bright to dark). The edge's thickness is proportional to the overlap weight. The node's size is proportional to its betweenness centrality. The node's color encodes the average year of the band in the data – from blue (low year, less recent) to red (high year, more recent).

strongly assortative by genre. With our division in four genres plus other, we observe an assortativity coefficient of 0.689, which is quite high. The assortativity coefficient for the average year of activity is even higher (0.91).

We omit showing the network using the regional information on the bands for two reason. First, there are too many regions (20) to visualize them by using different colors for nodes. Second, the structural relationship between the network and the regions is weaker – the assortativity coefficient being 0.223 – which would lead to a less clear visualization.

From the figures and the preliminary analysis, it appears quite evident that the structure of the network has a set of complex and interesting interactions with time, genres, and, to a lesser extent, geography. This means that it is meaningful to use the network structure to estimate the relationship between genres, time, and space. This is the main topic of the paper and we now turn our attention to this analysis.

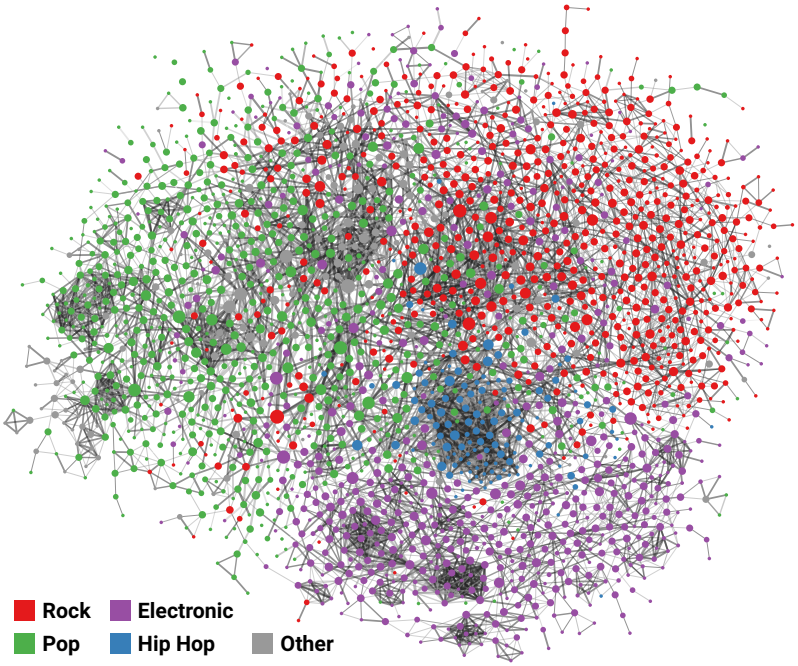



**Fig. 3**: The genre component of the band projection. Same legend as Figure 2, except for the node's color. Here, color encodes the dominant genre among pop (green), rock (red), electronic (purple), hip hop (blue), and other (gray).

# 3 Results

In this section we investigate a number of potential research questions in cultural data analytics. Each of them is tackled with a different node attribute analysis technique: network variance [1], network correlation [3, 60], and Generalized Euclidean distance [2] – which is at the basis of node attribute clustering [4] and era discovery. Supplementary Material Section 3 explains in details each of these methods.

## 3.1 Genre Specialization

When focusing on the genre attributes of the nodes, their network variance can tell us how concentrated or dispersed they are in the network. A disperse genre means that the bands playing that genre do not share artists, not even indirectly: they are scattered in the structure. Vice versa, a low-variance genre implies that there is a clique of artist playing it, and they are shared by most of the bands releasing records with that particular genre. Table 4 reports the five most (and least) concentrated genres in the network.

We only focus on genres that have a minimum level of use, in this specific case at least 1% of bands must have released at least one record using that specific genre. The values of network variance should be compared with a null version of the genre – the values themselves do not tell us whether they are significant or if we would get that

| Genre | Variance |
|---|---|
| Stoner Rock | 4.954** |
| Beat | 4.772*** |
| Neo-Classical | 4.605*** |
| Country | 4.403* |
| Post-Modern | 4.359* |
| ... | ... |
| Happy Hardcore | 0.249*** |
| Power Metal | 0.198*** |
| Eurobeat | 0.161*** |
| Gabber | 0.155*** |
| Trap | 0.105*** |

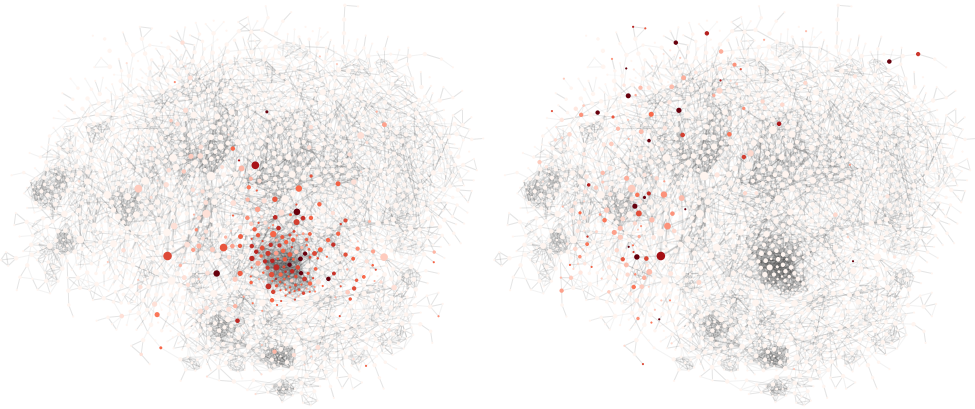
**Table 4**: The genres with the five highest and lowest variance in the band projection network. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

level of variance simply given the popularity of the genre. For this reason we bootstrap a pseudo p-value for the variance.

Let's assume that $\mathcal{S}$ is a $|V| \times |S|$ genre matrix. The $\mathcal{S}_{v,s}$ entry tells us how many records with genre $s$ the band $v$ has published. We can create $\mathcal{S}'$, a randomized null version of $\mathcal{S}$. In $\mathcal{S}'$, we ensure that each null genre has the same number of records as it has in $\mathcal{S}$. We do so by extracting with replacement at random $\sum_{v \in V} \mathcal{S}_{v,s}$ bands for genre $s$. The random extraction is not uniform: each band has a probability of being extracted proportional to $\sum_{s \in S} \mathcal{S}_{v,s}$. In this way, $\mathcal{S}'$ has the same column sum and similar row sum as $\mathcal{S}$. In other word, we randomize $\mathcal{S}$ preserving the popularity of each genre and each band. Then, we can count the number of such random $\mathcal{S}'$s in which the null genre has a higher (lower) variance than the observed genre.

Table 4 shows that stoner rock has a high and significant variance, indicating that bands playing stoner rock have a low degree of specialization. This can be contextualized by the fact that stoner rock was tried out unsystematically by a few unrelated bands, ranging from heavy metal to indie rock. On the other hand, many variants of heavy metal have low variance. This can be explained by the fact that heavy metal is a niche genre in Italy, and all bands playing specific heavy metal variants know each other and share members.

In Figure 4 we pick two representative genres – Hip Hop and Beat – which both have the same relatively high popularity in number of bands playing them, and have a significant (low or high) variance and we show how they look like on the network. The figure shows that the variance measure does what we intuitively think it should be doing: the Hip Hop bands have low variance and therefore strongly cluster in the network, while the Beat bands are more scattered.



(a) Hip Hop (b) Beat

**Fig. 4**: Two genres with different variance. Node size, node definition, and edge thickness, color, and definition is the same as Figure 2. The color is proportional to the genre-band node attribute value, with bright colors for low values and dark colors for high values.

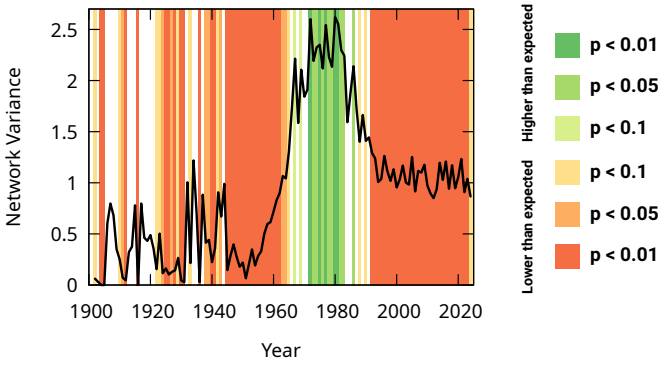



**Fig. 5**: The network variance (y axis) for a given decade (x axis). Background color indicates the statistical significance: red = lower than expected, green = higher than expected, white = not significantly different from expectation.

## 3.2 Temporal Variety

We are not limited to the calculation of variances for genres: we can perform the same operation for the years. If the variance of a genre tells us how diverse the set of bands playing is, the variance of a year can tell us how diverse the year was. Figure 5 shows the evolution of variances per year. We test the statistical significance of the observed variance value by shuffling the values of the node attribute for a given year a number of times, testing whether the observation is significantly higher, lower, or equal to this expectation.

From the figure we can see that there seems to be two phase transitions. In the first regime, we have an infancy phase with low activity and low variance. The first phase transition starts in the year 1960 and brings the network to a second regime of high activity and high variance. After the peak around the year 1980, a second phase transition introduces the third regime from the mid 90s until the present, with high activity but low variance. In the latter years, we see hip hop cannibalizing all genres and compressing the record releases to its tightly-knit cluster.

## 3.3 Node Attribute Correlation

We can now shift our attention from describing a single node attribute at a time – its variance as we saw in the previous sections – to describing the relationships between *pairs* of attributes. In this section, we do so by calculating their network correlation. Specifically, we want to make a geographical analysis. The ultimate aim is to answer the question: what are some particular strong genre-region associations? We can answer the question by calculating the network correlation between two node attributes, one recording the genre intensity for a band and the other a binary value telling us whether the band is from a specific region or not. The network correlation is useful here, because it grows not only if there are a lot of bands playing that specific genre in that specific region, but also if the other bands in the region that do not play that genre are close in the network to – i.e. share members with – bands playing that genre.

In Table 5 we report some significant region-genre associations. For each region, we pick the most popular genre in the network to which they correlate at a significant level – and they have the highest correlation among all other regions that correlate significantly to that genre. The significance is estimated via bootstrapping, by randomly shuffling the region vector – i.e. changing the set of bands associated to the region while respecting its size. Table 5 does not report a genre for all regions, because for some regions there was no genre satisfying the constraints. Note that some regions might correlate more strongly or more significantly with a genre that is not reported in the table, but we omit it if there was another region with a stronger correlation for that genre.

## 3.4 Genre Clusters

When we measure the pairwise distance between all node attributes systematically we can cluster them hierarchically. Here, we do such a network-based hierarchical clustering on the music genres and styles as recorded by Discogs. The aim is to see whether we can find groups of genres that are similar to each other, potentially informing a data-driven musical classification. Figure 6 shows a bird's eye view of the hierarchical clustering, with the similarity matrix and the dendrogram.

To make sense of it, we have selected some clusters, for illustrative purposes only. Table 6 shows what genres and styles from Discogs end up in the color-highlighted clusters from Figure 6. We can see that the clusters include similar genres which make as a coherent set of more general music styles. The figure also highlights that there is

| Region | Genre | Significance |
|---|---|---|
| Apulia | Latin | * |
| Calabria | Hip Hop | * |
| Campania | Folk, World, & Country | * |
| Emilia-Romagna | House | ** |
| Friuli | Reggae | *** |
| Lazio | Soundtrack | *** |
| Liguria | Prog Rock | *** |
| Lombardy | Electronic | *** |
| Marche | Rock | * |
| Piedmont | Punk | *** |
| Sardinia | Thrash | * |
| Sicily | Fado | ** |
| Tuscany | New Wave | * |
| Umbria | Cut-up/DJ | ** |
| Veneto | Krautrock | ** |

**Table 5**: The most popular genre with the strongest and significant correlation with a given region. The third column shows the significance level of the correlation, estimated via bootstrapping: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

| Color | Genres |
|---|---|
| Blue | Calypso, Bolero, Mambo, Cha-Cha, Tango, Beguine, Samba, Rumba |
| Green | Hard Rock, Symphonic Metal, Power Metal, Progressive Metal, Heavy Metal, Speed Metal, Thrash, Doom Metal, Death Metal, Black Metal |
| Purple | Trap, Pop Rap, Hip Hop, Conscious, Hardcore Hip-Hop, Boom Bap, Instrumental, Hip House |
| Black | Dub Techno, Acid, Deep Techno, Techno, Minimal, Tech House, Minimal Techno |

**Table 6**: The genres encased in the clusters we highlight in Figure 6.

a hierarchical structure of music styles, with meaningful clusters-within-clusters, and clear demarcation lines between groups and subgroups.

Recall that these clusters are driven exclusively by the network's topology and do not use any feature coming from the songs themselves. This means that using a network of shared members among bands is indeed insightful in figuring our the related genres these bands play. Therefore, network-based clustering has the potential to guide the definition of new musical classifications.

## 3.5 Temporal Clusters

We now look at the eras of Italian music we can discover in the data. Figure 7 shows the dendrogram, connecting years and groups of years at a small network distance to each other. Each era we identify colors its corresponding branch in the dendrogram. We avoid assigning an era for years pre-1906 and post-2018, due to issues with the representativeness of the data. We also notice that the 1938-1945 period is tumultuous, with many small eras in a handful of years, which is understandable given the geopolitical situation at a time, and so we ignore that period as well.

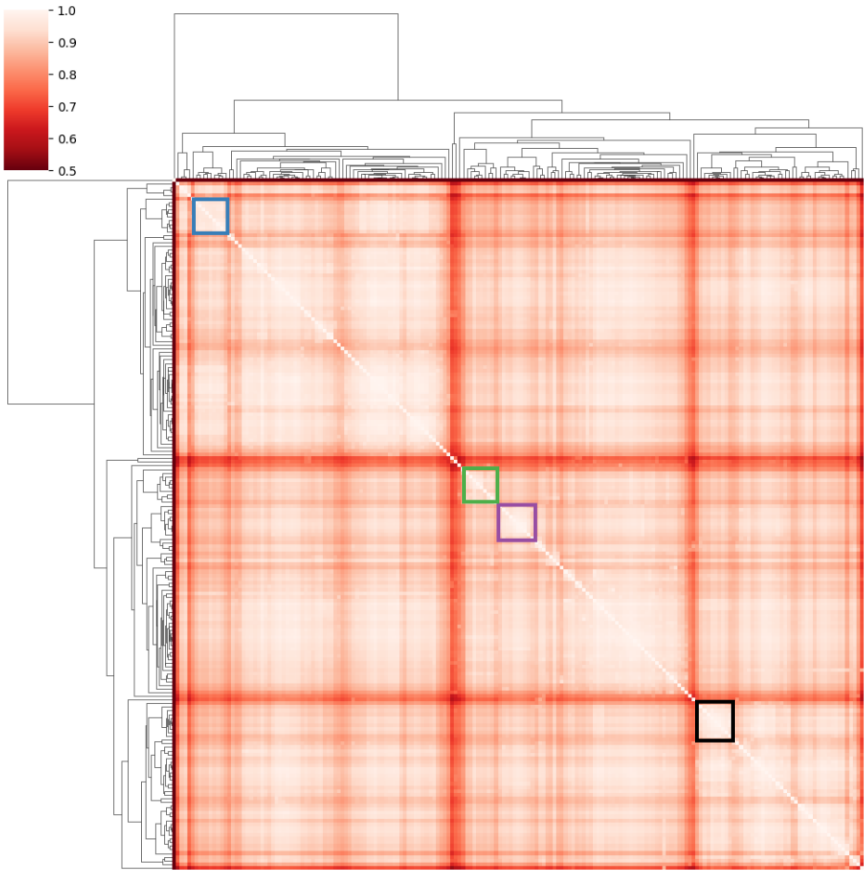



**Fig. 6**: The hierarchical genre clusters. The heatmap shows the pairwise similarity among the genres – from low (dark) to high (bright) similarity. The dendrograms show the hierarchical organization of the clusters.

To make sense of temporal clustering, the standard approach in the literature would be to compare counts of activities across clusters. However, that would ignore the role of the network structure. In our framework, we can characterize eras applying the same logic used to find them. We calculate the network distance between a node attribute representing the era and each genre. The era's node attribute simply reports, for each band, a normalized count of records they released within the bounds of that era. We normalize so that each era attribute sums to one, to avoid overpowering the signal with the scale of the largest and most active eras.

Then, for each era, we report the list of genres that have the smallest distance with that era. Note that some genres might still have a small distance with other eras, but we only report the smallest. These are the genres we use to label the eras in Figure 7. These genres are not the most dominant in that era – in almost all cases, pop and

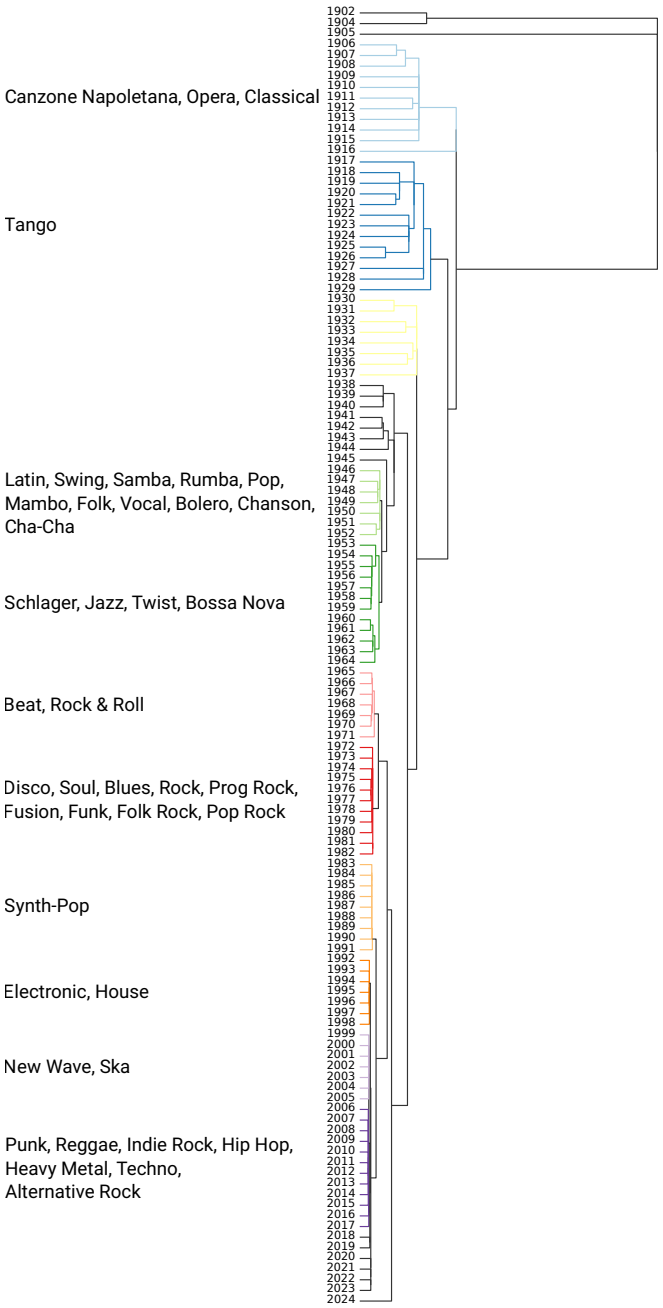



**Fig. 7**: The eras dendrogram. Clusters join at a height proportional to their similarity level (the more right, the less similar). Colors encode the detected eras with labels on the left.

rock dominate – but they give an intuition of what was the most characteristic genre of the era, distinguishing it from the others.

We can see that the characterization makes intuitive sense, with the classical genres being particularly correlated with the 1906-1916 era. Beat and rock'n'roll are particularly associated to the 1965-1971 period, the dates corresponding to the British Invasion in Italy. Notably, the punk genre has its closest association with the most recent era we label, 2006-2017, proving that – at least in Italy – punk is indeed not dead.

## 3.6 Explaining the Network

Wrapping up the analysis, one key assumption that underpins the analysis we made so far is that the connections in the band projection follow a few homophily rules. We can have meaningful genre (Section 3.4) and temporal (Section 3.5) clusters using our network distance measures only if bands do tend to connect if they have a genre or temporal similarity. Two bands should be more likely to share members if they play similar genres and if they do it at a similar point in time. More weakly, correlations between genres and geographical regions (Section 3.3) also make sense if bands with similar geographical origins also tend to share members more often than expected.

While proving this assumption would require a paper on its own, we can at least provide some evidence in favor of its reasonableness. We do so by running two linear regressions. In the first regression, we want to explain the likelihood of an edge to exist in the band projection with the genre, temporal, and geographical similarity between bands, or:

$$Y_{u,v} = \beta_0 + \beta_1 \mathcal{G}_{u,v} + \beta_2 \mathcal{R}_{u,v} + \beta_3 \mathcal{T}_{u,v} + \epsilon.$$

In this formula:

- $Y_{u,v}$ is a binary variable, equal to 1 if bands $u$ and $v$ shared at least one member, and zero otherwise;
- $\mathcal{G}_{u,v}$ is the genre similarity, which is the cosine similarity between the vectors recording how many records of a given genre bands $u$ and $v$ have published;
- $\mathcal{R}_{u,v}$ is the region similarity, equal to 1 if the bands originate from the same region, and zero otherwise;
- $\mathcal{T}_{u,v}$ is the temporal similarity, in which we take the logarithm of the number of years in which both bands released a record, plus one to counter the issue when the bands did not share a year;
- $\beta_0$ and $\epsilon$ are the intercept and the residuals.

Note that $Y_{u,v}$ contains all links with weight of at least one, even those that are not statistically significant and were dropped from our visualizations and analyses from the previous sections. Moreover, it also has to contain all non-links. However, since the network is sparse, it is not feasible to have all non-links in the regression. Thus, we perform a balanced negative sampling: for each link that exists we sample and include in $Y_{u,v}$ a link that does not.

For $\mathcal{G}_{u,v}$ we only consider the most popular 38 genres, since sparsely used genres would make bands more similar than what they would otherwise be.

| | *Dependent variable:* | |
|---|---|---|
| | Exists | Size |
| | (1) | (2) |
| Genre | $0.568^{***}$ | $0.605^{***}$ |
| | (0.004) | (0.008) |
| Region | $0.079^{***}$ | $0.089^{***}$ |
| | (0.003) | (0.006) |
| Year | $0.248^{***}$ | $0.325^{***}$ |
| | (0.001) | (0.003) |
| Constant | $0.210^{***}$ | $0.037^{***}$ |
| | (0.002) | (0.005) |
| Observations | 173,966 | 86,983 |
| $R^2$ | 0.284 | 0.170 |
| Adjusted $R^2$ | 0.284 | 0.170 |
| Residual Std. Error | 0.423 (df = 173962) | 0.693 (df = 86979) |
| F Statistic | $22{,}966.210^{***}$ (df = 3; 173962) | $5{,}944.348^{***}$ (df = 3; 86979) |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

**Table 7**: The regression results from our two models predicting the existence of a link (column 1) and its weight (column 2).

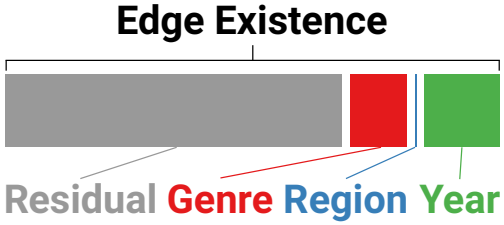



**Fig. 8**: The relative importance of each explanatory variable to determine the existence of a link between two bands in the band projection.

The first column of Table 7 shows the result of the model. The first thing we can see is that we can explain 28.4% of the variance in the likelihood of a edge to exist. This means that 71.6% of the reasons why two bands share a member is not in our data – be it unrecorded social networks, random chance, impositions from labels, etc.

However, explaining 28.4% of the variance in the edge existence likelihood still provides a valid clue that our homophily assumptions should hold. All similarities we considered play a role in determining the existence of an edge: all of their coefficients are positive and statistically significant. Given that these similarity measures do not share the same units – and not even the same domain –, one cannot compare the coefficients directly. However, we can map their contributions to the $R^2$ by estimating their relative importance [61, 62], which we do in Figure 8. From the figure we can see that it is the temporal similarity the one playing the strongest role, closely followed by genre similarity. Spatial similarity, on the other hand, while still being statistically significant, provides little to no additional explanatory power to the other factors.

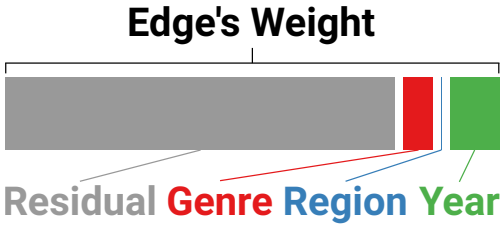



**Fig. 9**: The relative importance of each explanatory variable to determine the weight of a link between two bands in the band projection.

Once we establish that the *existence* of the connection is related to genre, temporal, and geographical similarity, we can ask the same question about the *strength* of the relationship between two bands. We apply the same model as before, changing the target variable:

$$\log(W_{u,v}) = \beta_0 + \beta_1 \mathcal{G}_{u,v} + \beta_2 \mathcal{R}_{u,v} + \beta_3 \mathcal{T}_{u,v} + \epsilon.$$

Here, $\log(W_{u,v})$ is the logarithm of the edge weight. Note that here we only focus on those edges that have a non-zero weight, i.e. those that exist. This is because we do not want this model to try and predict also edge existence, beside its strength, as we already took care of that problem with the previous model.

Table 7 contains the results in its second column. We can see that, also in this case, all three factors are significant predictors of the edge weights. The number of artists two bands share goes up if the two bands play similar genres, with temporal overlap, and if they originate from the same region. The $R^2$ is noticeably lower, though, which means that $\log(W_{u,v})$ is harder to predict than $Y_{u,v}$.

Figure 9 shows the same $R^2$ decomposition we did in Figure 8 for $Y_{u,v}$. All explanatory variables explain less variance than in the previous model. Relative to each other, the temporal overlap is the factor gaining more importance than genre similarity.

# 4 Discussion

In this paper we have provided a showcase of the analyses and conclusions one could do in cultural data analytics by using node attribute analysis. We focused on the case study of Italian music from the past 120 years. We built a bipartite network connecting artists to bands and then projected it to analyze a band-band network. We have shown how one could identify genres concentrating in such a network, hinting at clusters of bands playing homogeneous genres, using network variance. We have shown a geographical analysis, calculating the network correlation between the region of origin of bands and the genres they play. We have shown how one could create a new music genre taxonomy by performing node attribute clustering on music genre data. We also proposed a novel way of performing era detection in a network, by finding clusters of similar consecutive years, where years are node attributes.

While we believe our analysis is insightful, there are a number of considerations that need to be made to contextualize our work. We can broadly categorize the limitations

in two categories: the one relating to the domain of analysis, and the methodological ones.

When it comes to cultural data analytics, we acknowledge the fact that we are working with user-generated data. There is no guarantee that the data is free from significant mistakes, misleading entries, and incompleteness. Furthermore, our results might not be conclusive. We process data semi-automatically, and the coding process is not complete, meaning we miss a considerable amount of the lesser known artists. This also means that there could be biases in the data collection, induced by our decision on the order in which we explore the structure – which might be focusing too much or too little on specific areas of Italian music. As a specific example, in our project we have ignored another potentially rich source of node attributes: information about the music labels/publishers. This is available on Discogs, and we could envision a label to be represented as a node vector, whose entries are the number of records a specific label published for a specific band. We plan to use this information for future work. The coding process is still ongoing, and we expect to be able to complete the network in the near future.

On the methodlogical side, we point out that what we did is only possible in the presence of rich metadata – dozens if not hundreds of node attributes. Networks with scarce node attribute data would not be amenable to be analyzed with the techniques we propose here. However, in cultural data analytics, there is usually a high richness of metadata. Furthermore, many of the node attribute techniques only make sense if the node attributes are somehow correlated with the network structure. The musical genre clustering or the era detection would not produce meaningful results if the probability of two nodes of connecting was not influenced by their attributes – i.e. if the homophily hypothesis does not hold. In our case, the homophily assumption likely holds, as we show in Section 3.6.

When considering some specific analyses we performed other limitations emerge. For instance, our era discovery approach exclusively looks at node activities. However, structural changes in the network's connections also play a key role in determining discontinuities with the past [63]. We should explore in future work how to integrate our node attribute approach with structural methods. When it comes to the use of network variance, how to properly estimate its confidence intervals without using bootstrapping remains a future work. Therefore, the results we present here should be taken with caution, as it might be that some of the patterns we highlight are not statistically significant.

On a more practical side, our node attribute techniques hinge on specific matrix operations. While these can be efficiently computed on GPU using tensor representations, this might put a limit on the size of the networks analyzed, which have to fit in the GPU's memory.

## Declarations

**Availability of data and materials**. All data and code necessary to replicate our results are available at http://www.michelecoscia.com/?page_id=2336 and [55].

**Competing interests**. The authors declare that they have no competing interests.

**Funding**. No funding was provided for this paper.

**Authors' contributions**. M.C. designed and performed all experiments, prepared figures, and wrote and approved the manuscript.

**Acknowledgements**. The author is thankful to Amy Ruskin for the project's idea, and to Seth Pate and Clara Vandeweerdt for insightful discussions.

# References

[1] Devriendt, K., Martin-Gutierrez, S., Lambiotte, R.: Variance and covariance of distributions on graphs. SIAM Review **64**(2), 343–359 (2022)

[2] Coscia, M.: Generalized euclidean measure to estimate network distances. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, pp. 119–129 (2020)

[3] Coscia, M.: Pearson correlations on complex networks. Journal of Complex Networks **9**(6), 036 (2021)

[4] Damstrup, A.S.R., Madsen, S.T., Coscia, M.: Unsupervised learning via network-aware embeddings. arXiv preprint arXiv:2309.10408 (2023)

[5] Manovich, L.: Cultural Analytics. Mit Press, ??? (2020)

[6] Candia, C., Jara-Figueroa, C., Rodriguez-Sickert, C., Barabási, A.-L., Hidalgo, C.A.: The universal decay of collective memory and attention. Nature human behaviour **3**(1), 82–91 (2019)

[7] Schich, M., Hidalgo, C., Lehmann, S., Park, J.: The network of subject co-popularity in classical archaeology (2008)

[8] Brughmans, T.: Thinking through networks: a review of formal network methods in archaeology. Journal of archaeological method and theory **20**, 623–662 (2013)

[9] Mills, B.J., Clark, J.J., Peeples, M.A., Haas Jr, W.R., Roberts Jr, J.M., Hill, J.B., Huntley, D.L., Borck, L., Breiger, R.L., Clauset, A., *et al.*: Transformation of social networks in the late pre-hispanic us southwest. Proceedings of the National Academy of Sciences **110**(15), 5785–5790 (2013)

[10] Salah, A.A., Manovich, L., Salah, A.A., Chow, J.: Combining cultural analytics and networks analysis: Studying a social network site with user-generated content. Journal of Broadcasting & Electronic Media **57**(3), 409–426 (2013)

[11] Hristova, S.: Images as data: cultural analytics and aby warburg's mnemosyne. International Journal for Digital Art History (2) (2016)

[12] Karjus, A., Solà, M.C., Ohm, T., Ahnert, S.E., Schich, M.: Compression ensembles quantify aesthetic complexity and the evolution of visual art. EPJ Data Science

**12**(1), 21 (2023)

[13] Bail, C.A.: The cultural environment: Measuring culture with big data. Theory and Society **43**, 465–482 (2014)

[14] Hohmann, M., Devriendt, K., Coscia, M.: Quantifying ideological polarization on a network using generalized euclidean distance. Science Advances **9**(9), 2044 (2023)

[15] Ronen, S., Gonçalves, B., Hu, K.Z., Vespignani, A., Pinker, S., Hidalgo, C.A.: Links that speak: The global language network and its association with global fame. Proceedings of the National Academy of Sciences **111**(52), 5616–5622 (2014)

[16] McAndrew, S., Everett, M.: Music as collective invention: A social network analysis of composers. Cultural Sociology **9**(1), 56–80 (2015)

[17] Vlegels, J., Lievens, J.: Music classification, genres, and taste patterns: A ground-up network analysis on the clustering of artist preferences. Poetics **60**, 76–89 (2017)

[18] Moon, S.-I., Barnett, G.A., Lim, Y.S.: The structure of international music flows using network analysis. New Media & Society **12**(3), 379–399 (2010)

[19] Stebbins, R.A.: Music among friends: the social networks of amateur musicians'. Popular music: Critical concepts in media and cultural studies **1**, 227–245 (2004)

[20] Park, J., Celma, O., Koppenberger, M., Cano, P., Buldú, J.M.: The social network of contemporary popular musicians. International Journal of Bifurcation and Chaos **17**(07), 2281–2288 (2007)

[21] Gleiser, P.M., Danon, L.: Community structure in jazz. Advances in complex systems **6**(04), 565–573 (2003)

[22] Teitelbaum, T., Balenzuela, P., Cano, P., Buldú, J.M.: Community structures and role detection in music networks. Chaos: An Interdisciplinary Journal of Nonlinear Science **18**(4) (2008)

[23] Baym, N.K., Ledbetter, A.: Tunes that bind? predicting friendship strength in a music-based social network. Information, Communication & Society **12**(3), 408–427 (2009)

[24] Pennacchioli, D., Rossetti, G., Pappalardo, L., Pedreschi, D., Giannotti, F., Coscia, M.: The three dimensions of social prominence. In: International Conference on Social Informatics, pp. 319–332 (2013). Springer

[25] Pálovics, R., Benczúr, A.A.: Temporal influence over the last. fm social network. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances

in Social Networks Analysis and Mining, pp. 486–493 (2013)

[26] Aljalbout, E., Golkov, V., Siddiqui, Y., Strobel, M., Cremers, D.: Clustering with deep learning: Taxonomy and new methods. arXiv preprint arXiv:1801.07648 (2018)

[27] Aggarwal, C.C., *et al.*: Neural networks and deep learning. Springer **10**(978), 3 (2018)

[28] Pang, G., Shen, C., Cao, L., Hengel, A.V.D.: Deep learning for anomaly detection: A review. ACM computing surveys (CSUR) **54**(2), 1–38 (2021)

[29] Ezugwu, A.E., Ikotun, A.M., Oyelade, O.O., Abualigah, L., Agushaka, J.O., Eke, C.I., Akinyelu, A.A.: A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Engineering Applications of Artificial Intelligence **110**, 104743 (2022)

[30] Mahalanobis, P.: On the generalized distance in statistics. (1936). National Institute of Science of India

[31] Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: International Conference on Machine Learning, pp. 478–487 (2016). PMLR

[32] Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P.: Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine **34**(4), 18–42 (2017)

[33] Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE transactions on neural networks **20**(1), 61–80 (2008)

[34] Wu, L., Cui, P., Pei, J., Zhao, L., Guo, X.: Graph neural networks: foundation, frontiers and applications. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 4840–4841 (2022)

[35] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph neural networks: A review of methods and applications. AI open **1**, 57–81 (2020)

[36] Bo, D., Wang, X., Shi, C., Zhu, M., Lu, E., Cui, P.: Structural deep clustering network. In: Proceedings of the Web Conference 2020, pp. 1400–1410 (2020)

[37] Tsitsulin, A., Palowitch, J., Perozzi, B., Müller, E.: Graph clustering with graph neural networks. arXiv preprint arXiv:2006.16904 (2020)

[38] Bianchi, F.M., Grattarola, D., Alippi, C.: Spectral clustering with graph neural networks for graph pooling. In: International Conference on Machine Learning, pp. 874–883 (2020). PMLR

[39] Zhou, K., Huang, X., Li, Y., Zha, D., Chen, R., Hu, X.: Towards deeper graph neural networks with differentiable group normalization. Advances in neural information processing systems **33**, 4917–4928 (2020)

[40] Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: Online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 701–710 (2014)

[41] Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. Advances in neural information processing systems **30** (2017)

[42] Perozzi, B., Akoglu, L., Iglesias Sánchez, P., Müller, E.: Focused clustering and outlier detection in large attributed graphs. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1346–1355 (2014)

[43] Zhang, C., Song, D., Huang, C., Swami, A., Chawla, N.V.: Heterogeneous graph neural network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 793–803 (2019)

[44] Wang, C., Pan, S., Hu, R., Long, G., Jiang, J., Zhang, C.: Attributed graph clustering: A deep attentional embedding approach. arXiv preprint arXiv:1906.06532 (2019)

[45] Lin, Z., Kang, Z., Zhang, L., Tian, L.: Multi-view attributed graph clustering. IEEE Transactions on knowledge and data engineering (2021)

[46] Cheng, J., Wang, Q., Tao, Z., Xie, D., Gao, Q.: Multi-view attribute graph convolution networks for clustering. In: Proceedings of the Twenty-ninth International Conference on International Joint Conferences on Artificial Intelligence, pp. 2973–2979 (2021)

[47] Yang, S., Verma, S., Cai, B., Jiang, J., Yu, K., Chen, F., Yu, S.: Variational co-embedding learning for attributed network clustering. Knowledge-Based Systems **270**, 110530 (2023)

[48] Bretto, A.: Hypergraph theory. An introduction. Mathematical Engineering. Cham: Springer **1** (2013)

[49] Bianconi, G.: Higher-order Networks. Cambridge University Press, ??? (2021)

[50] Benson, A.R., Gleich, D.F., Leskovec, J.: Higher-order organization of complex networks. Science **353**(6295), 163–166 (2016)

[51] Lambiotte, R., Rosvall, M., Scholtes, I.: From networks to optimal higher-order models of complex systems. Nature physics **15**(4), 313–320 (2019)

[52] Xu, J., Wickramarathne, T.L., Chawla, N.V.: Representing higher-order dependencies in networks. Science advances **2**(5), 1600028 (2016)

[53] Kaufman, T., Oppenheim, I.: High order random walks: Beyond spectral gap. Combinatorica **40**, 245–281 (2020)

[54] Carletti, T., Battiston, F., Cencetti, G., Fanelli, D.: Random walks on hypergraphs. Physical review E **101**(2), 022308 (2020)

[55] Coscia, M.: Italian XX-XXI Century Music. https://doi.org/10.5281/zenodo.13309793 . https://doi.org/10.5281/zenodo.13309793

[56] Newman, M.E.: Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. Physical review E **64**(1), 016132 (2001)

[57] Zhou, T., Ren, J., Medo, M., Zhang, Y.-C.: Bipartite network projection and personal recommendation. Physical Review E **76**(4), 046115 (2007)

[58] Yildirim, M.A., Coscia, M.: Using random walks to generate associations between objects. PloS one **9**(8), 104813 (2014)

[59] Coscia, M., Neffke, F.M.: Network backboning with noisy data. In: 2017 IEEE 33rd International Conference on Data Engineering (ICDE), pp. 425–436 (2017). IEEE

[60] Coscia, M., Devriendt, K.: Pearson correlations on networks: Corrigendum. arXiv preprint arXiv:2402.09489 (2024)

[61] Feldman, B.E.: Relative importance and value. Available at SSRN 2255827 (2005)

[62] Grömping, U.: Relative importance for linear regression in r: the package relaimpo. Journal of statistical software **17**, 1–27 (2007)

[63] Berlingerio, M., Coscia, M., Giannotti, F., Monreale, A., Pedreschi, D.: Evolving networks: Eras and turning points. Intelligent Data Analysis **17**(1), 27–48 (2013)