An Empirical Study of Why You Should Support Your Local Bands

Candidate 48743

August 14, 2025

Word Count: 9844

Abstract

This paper constructs and analyses a dataset of the geography of recorded music. By geocoding information from a repository of digitised liner notes, I construct a novel database of the recording locations of over 270,000 albums. I find that large urban areas are significantly less likely to produce albums that are forerunners of trends. However, there is evidence that large urban areas appear to produce the most commercially successful albums. Constructing a graph of collaborations within cities using personnel information, I find that highly innovative cities are associated with higher density networks of collaboration.

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Contents

1 Introduction

Economists and social scientists who study cities tend to extol the virtues of agglomeration as a means to turbocharge the creative process by bringing together individuals with ideas, technical know-how, and capital. The urban world's strength in innovating through collective learning is attributed first to Alfred Marshall who identified geographic proximity in knowledge diffusion as a source for agglomeration. Theorists like Jane Jacobs later refined this notion, arguing that the key to urban success is in diversity and creativity: new ideas emerge in clusters where differences in industry, methodology, and ideology are forced together [FAM17].

Many attempts have been made by economists to quantify the extent of the relationship between cities and creativity. This line of literature is overwhelmingly focused on technological innovation, going so far as to use creativity and innovation as near synonyms [Cas]. Generally, it identifies urban size as the key driver of innovation cross space [CK]. However, the overwhelming focus on technological innovation leaves a major gap for quantifying how these forces manifest in other spheres of creative life. Art, literature, music, and ideologies all have geographic patterns of production that mirror those of conventional industries. A few papers have described the geography of creative industries [BDFH15] [BHM], using employment shares and wage data to argue that creative industries face the same agglomeration forces as conventional industries. Currently no significant literature attempts to describe the spatial distribution of creative industries using information about the creative product itself.

We have good reason to believe that creativity in the arts may behave fundamentally differently to creativity in technological innovation with regard to spatial distribution. Barriers to its production are lower and the artists are more likely than inventors to benefit from disseminating their work. Experimentation can occur with far less cost. Nevertheless, the availability of data on technological innovation has produced a wealth of methodologies for studying the role of cities in creativity. This paper aims to fill this gap by building a novel dataset of music production across time and space, with rich enough information to produce a detailed account of where and how novel forms of music making emerge.

Discogs [dis25] is a large online marketplace for physical media and is generally considered the benchmark for pricing albums [CS20]. In addition to a marketplace function, Discogs acts as a musical archive, storing a large amount of metadata attached to every album including the album's personnel, copyright holders, tracklist, genre tags, and recording information. The key insight in this paper's construction is connecting the recording studios listed in an albums liner notes to the contact information listed on the studio database. This contact page contains enough information to geocode a large fraction of albums in the data. Being able to encode locations at a large scale is crucial, as past research that samples from only charting or well known albums is unlikely to be reflective of the overall picture of the industry [GKB21]. From this database of albums we construct a comprehensive innovation score for each album by describing the time-sensitive semantic distinctiveness of a set of qualitative descriptors the album.

This paper studies the extent to which artistic innovation is driven by agglomeration. First, it examines agglomeration effects in musical creativity. I find that large cities do not create greater innovation with regard to the creation of new subgenres and combinations of musical styles. On the contrary, I find robust evidence that smaller cities in our data are ahead of trends relative to large cities. Second, I find an ambiguous relationship between a city's specialization in particular genres and its creative potential. Using the movement of musicians as an indicator of connectedness between cities, I find that cities central to the network of artists produce albums that are more consistent with trends but less innovative. Third, I examine characteristics of collaboration networks within cities where I identify that dense networks tend to produce greater stylistic innovation. This evidence is robust to the same tests conducted on our first hypothesis.

This paper introduces several improvements over previous research in the field. It is, to my knowledge, the first method geocode music at the point of production at a large scale. Previous work has relied on the birthplaces of artists [Mak17], listeners [LMC13], or concerts [Li] and generally lacks the ability to look across urban areas and musical styles. Second, it is the first study to attempt to measure the innovative-ness of an album using metadata and the first to introduce a measure of innovation that can theoretically be analysed across genres.

2 Prior Research

2.1 Conventional Measures of the Geography of Innovation

Innovation is a critical area of research in economic geography. Much of the success of superstar cities in the past three decades has been attributed to their superior size and diversity. Duranton and Puga (2004) [DP01]provide the most widely cited account of how dense, large urban areas serve as hotbeds of innovation. Their model describes "nursery cities": Early industries benefit strongly from inter-industry knowledge spillovers. Over time, technologically mature industries benefit less from proximity and disperse to more affordable labour markets or smaller, specialized clusters. While many papers paint this as the consensus view, contradictions and counter examples exist. Fritsch and Wyrwich's (2021) [FW] case study of Germany's post-war innovative geography suggests that institutional factors may overwhelm the natural tendency of large cities to dominate in innovation. Other commentators point to Silicon Valley: today a major hub of tech companies, at the time of its emergence the areas that today comprise this cluster were unremarkable suburbs.

The workhorse method of studying creativity and the emergence of ideas in cities is through patent citations. This line of literature arose from Jaffe, Trajtenberg, and Henderson (1993) [JTH93], which identified patent citations as a way to map the knowledge transition process to particular outputs. Since this insight there has been a deluge of empirical research into the characteristics of patents in various cities or regions. Patents have a trifecta of properties that make them extremely useful for identifying innovation in cities: they can be located with reasonable precision, they have multiple objective metrics of usefulness and quality (see Kogan et. al [KPSS12]), and through citations they have a mechanism to tell us about the origin of knowledge that led to the invention. Such data can be accessed from large, well-maintained repositories such as PATSTAT. Inventors can be easily identified, as can the corresponding firms and their R&D spending levels.

Berkes and Gaetani (2021) [BG21]study novelty in patenting, identifying the propensity of larger and denser cities to produce "unconventional" inventions. To measure unconventionality, they consider the expected distribution of technology codes of citing patents. By identifying patents that cite from an unexpected combination of technologies the authors are able to determine the geographic distribution of these unconventional patents. They find that these novel combinations are far more likely than in the general case to be produced in large, economically diversified cities. The key weakness of this paper, however, is that they do not control for the impact of patents. While legal considerations are generally eschewed by economics literature on patents, it is possible that these novel patents arise in large firms to create "patent thickets" designed to hinder potential future innovation in these respective fields.

Castaldi (2023) [Cas] gives a recent attempt to expand the scope of data in analysing creativity in cities. Her core insight is to integrate design rights and trademarks, two additional forms of intellectual property. While her findings largely reinforce the existing patent research, the results suggest that smaller cities are more innovative than a pure focus on patents suggests.

Outside of patent research, academic citation networks also provide and means to study the effects of agglomeration of novel idea formation. Nomaler et al (2014) [NFH14] find evidence of agglomeration's effect on per-capita academic output. Larger cities produce higher per capita scientific research, though the authors note that the largest cities in the data deviate from the trend, suggesting that the effects of agglomeration on knowledge creation weaken at the very top.

The difficulty in this line of research is isolating the effects of agglomeration on creativity in a strict sense. The success of innovation in places like Silicon valley is related to factors beyond the transmission of ideas including access to capital or a customer based for business to business sales. In describing Jacobian externalities using resource intensive knowledge production outputs like patents or academic research, we do not properly isolate the effect of agglomeration on the types of creative forces that may power other aspects of culture.

2.2 The Geography of Music

Empirical work on the geography of music mostly relies on measuring production inputs rather than outputs. Labour market inputs - workers in the music industry - can be easily tracked using conventional data sources. Florida and Jackson 2010 [FJ10] document the shifting patterns in music's economic geography by using Bureau of Labour Statistics data on musical employment. They find increasing concentration in employment and establishment density between 1970 and 2004. A report on the UK's creative economy from Nesta [BDFH15] provides a more granular picture of employment. They find that employment, establishments, and investment are overwhelmingly concentrated in the South East, particularly in London.

Verboord and Noord (2016) [VvN16] study the empirics of music scenes in the context of studying how the social media affects the relative strengths of large and small cities by removing geographic proximity to traditional gatekeepers. The crucial issue in this study is the fact that their dataset scrapes only the current location of artists to classify. The confirm a high degree of spacial concentration of the music industry but suggest that this has been decreasing over time, contrasting with the evidence presented by Florida and Jackson (2010) [FJ10]

A more precise approach to musical geography, and one more in keeping with the data mining approach of this paper is Lee and Cunnningham (2012) [LMC13] who examine the geography of novel music consumption rather than production. Using data from Last.fm, a social network that gathers data on its users listening habits from streaming services, the authors determine leaders and laggards in new trends using geolocated data on the listeners. The authors find only weak evidence suggesting that large cities are consistently ahead of trends. The main focus of the paper is thus on identifying consistent patterns of directionality in trends. Li 2023 [Li] uses Spotify data to geocode over 100,000 country music concerts across the contiguous United States. Watson's (2012) [Wat12] bears the most similarity to this paper. It is, to my knowledge, the only other paper that encodes the location of music using the recording locations of albums. Watson identifies the studios using iTunes metadata and geocodes the studios by hand. The result enables the creation of networks of musical collaboration between cities by looking at albums with recording locations in multiple cities. Watson's sample is not representative of the universe of music - sampling from the iTunes album charts during an six month period in 2009. It also produces three different graphs for charting albums in the US, UK and Australia. These factors limit the validity of the paper's descriptive analysis.

Empirical research in music lacks scale. Where scale is achieved, insights beyond descriptive analysis are sparse. By and large, studies of the formation, categorization, and success of local music scenes remain qualitative. Most research is purely desciptive of a particular city or trend. Azzerad (2001) [Aze01] chronicles the development of early indie rock music in the 80s from its origin in the Hardcore Punk movement to the alternative rock and grunge that emerged in the early 90s. While not explicitly focussed on geography, Azzerad's thorough reporting presents a view of creative geography distinct from the agglomeration approach taken by much of the previous literature. The scenes Azzerad identifies as influential are marginal to the mass of recorded music at the time, notably the South Bay region of California, Minneapolis, and Washington D.C. Creativity here is still deeply social, but the structures that induce creativity are dense networks operating at a small scale. While isolated to a particular case study and devoid of empirical research, Azzerad's view of creativity is one large vindicated by the results of this paper. In a similar vein, Cohen 1991 [Coh91] profiles the musical history of Liverpool throughout two booms in the city's musical heritage: the early 60s and the early 80s. While Cohen identifies factors that would be familiar to urban economists such as Liverpool's port activities drawing in diverse global influences, she also paints a view of the cities bands as scrappy and small scale. She emphasises the large extent of shared social networks and churn in group membership as a key element in the diffusion of ideas across musicians. Thus the literature is split. Empirical work, currently far from sufficient, suggests that music is concentrated by the same forces that concentrate high technology firms. However, qualitative research paints a more nuanced picture.

2.3 Network Determinants of Innovation

Viewing creativity as an inherently social process, arguably the most compelling research focusses on social network structures as determinants of innovative activity. Collaboration networks in music are much more well studied. Discogs' ability to construct networks accurately is already well known among researchers with open source implementations for network generation [Obend][1]. The most comprehensive treatment of graph descriptions of music comes from Gienapp et. al (2018) [GKB21] who examine the universe of Discogs data (as of 2018) of Hip Hop and Jazz music.

An attempt at using musical network data to assess creative potential comes from Budner and Grahl (2016) [BG16]. Drawing from the same data source as this paper, their research builds a collaboration

[1] The implementation of graph construction described in in the section 3 is entirely the author's

network based on the artists affiliated with Rolling Stone Magazine's list of the 500 "greatest albums of all time". While their work discusses properties of this particular network, there are several barriers holding it back from being a meaningful assessment of collaboration. Firstly, there is no comparative analysis to distinguish properties of musical networks generally from those of particularly successful or innovative networks. Second, isolating the network to only well known releases substantially limits its explanatory power. Connections though collaboration on external projects, or collaborations early in the artists career are likely to be lost.

Beyond music, significant research has focussed on characteristics of social networks in innovation, again using patents are a measure of innovation. Unlike research on general agglomeration effects, studies of networks in innovation are generally smaller scale; employee transitions between firms and within-firm hierarchies are difficult to find data for. Guler and Nekar 2010 [GN12] identify a particularly interesting pattern in the innovation response to network density. Their study concerns patenting patterns amongst pharmaceutical companies in the 1980s. They draw a distinction between 'local' and 'global' network interconnectedness. Density of networks within research teams and firms are found to be correlated with innovation, consistent with the social hypothesis of creativity. However, density a global level - indicative of collaboration between firms - is negatively associated with innovation. Though their sample only consists of 33 firms across a single decade, their results largely replicate, notably in Li and Zhou's (2024) [ZL] study of 2,337 Chinese R&D intensive biochemical firms.

Breschi and Lenzi 2016 [BL] examine network density of innovators using coauthorship on patents. They first compute network density within the "clique" of inventors in a given city. They then compute a second network for each inventor, considering only edges in the graph corresponding to connections outside the inventors urban area. Largely in keeping with the studies previously mentioned, they find a significant positive effect on innovation from network density, but only from the network of individuals within the city. This line of research generates two hypotheses concerning density:

1. Density within small-scale networks causes greater innovation [BL].
2. Interconnectedness at a global level either does not cause innovation [BL] or actively harms innovation [GN12].

3 Data Construction

3.1 Discogs

The data in this paper is almost entirely drawn from Discogs.com. Discogs is an online repository of music consisting of the digitised liner notes of 17 million albums, singles, live recordings, and compilations. The database originally intended to catalogue physical media but has expanded in recent years to include digital formats, though digital-only music is poorly represented [ANNN21]. The database includes conventional albums alongside live recordings, compilation albums, singles, and EPs. Discogs releases monthly data dumps; this paper uses the data released on 1st April, 2025 [dis25]. The data consists of XML files each representing four key data types: releases, masters, artists, and labels.

The releases file contains records concerning a particular "release" of an album, single, compilation, mixtape, or live recording. Albums, especially particularly popular ones, have multiple releases coinciding with different products. Most commonly, albums will have separate releases for different countries or separate releases for different formats.

Take, for example, the album *Land Speed Record* by Hüsker Dü. It has 20 releases in Discogs. It was originally released in the US and the UK under separate labels in 1982, constituting two releases in that year. Two more LP releases occurred under the original label in 1983 and 1984, and a cassette was issued in 1986, taking the total releases to five. After the band joined hardcore punk label SST in the mid-80s, the album was reissued several times in the years since, totalling 15 more separate releases. Separate releases can be attributed to differences as trivial as alternative sleeves or the color of the vinyl used in pressing — Taylor Swift's album *Midnights* has 11 LP releases in 2022 on account of the variety of colours and special editions used to market it.

The differences between releases from the same album are not important. However, since the liner notes can vary between editions of the album, the vast majority of important information to be extracted from this data sits at this level of analysis. Crucially, the release xml contains information on the recording studios used in the project and the full list of collaborators listed in the album's liner notes.

Releases that describe the same album are linked to a "master release" in the 'masters' database. Each master is a separate musical project and forms the basic unit of analysis in this paper. Not every release has a master, but every master has at least two releases. While I use information from the release level, I only consider releases that map to a master, which removes most singles and one-off releases. When aggregating information from the release to master level I make use of Discogs quality control flag. Where the flag "Correct and Complete" is present in a release, I import that release information as the full set of master metadata. Where this is not, present I merely take all listed studios or collaborators from any release. I assume that differences in the release metadata from the same albums reflects adding or removing information from the liner notes across versions.

Another critical piece of information in the masters database is genre metadata. Albums on Discogs contain both "genre" tags and "style" tags. There are 14 genre tags and 686 possible style tags. Releases uploaded to the database are required to carry these tags. Thus these tags are present in 100% of releases, though the qualitative nature of the tags make it hard to assess their accuracy. A style is not a strict subtype of a given genre. Style can include both qualitative descriptors of music such as "ambient", "abstract", or "downtempo" as well as more specific subgenres like "Speed Metal", "Garage Rock", or "Donk". Despite both "genre" and "style" indicating information about the qualitative nature of the music, this paper uses the two tag sets in completely different ways. In this study, genre is used to divide the albums, describe the diversity of music within a given city, and control for unobserved differences between genres. The more fluid and granular style tags are primarily used to construct our measurement of innovation, discussed more in section 3.6.

The number of releases per master is cautiously used as a measure of the commercial success of an album. Distribution to more countries or releases in more formats is likely indicative of larger demand for that record. While metrics such as streams or chart performance may be more accurate, releases per master is the best way to capture the performance of records that would otherwise not register on charts in a way that is consistent across our data. Among the problematic reasons for a large number of releases, qualities such as alternative cover art or colors are unlikely to produce bias when aggregated up to the city level. Two noteworthy factors affecting releases per master need to be controlled for. Older albums have more re-releases and certain time periods may have a greater variety in formats that necessitates more releases to reach the same audience. Geographically, the data suggest that smaller markets have more releases per master - reaching the same size audience from a European country may necessitate releases in more countries than in the US. Indeed, at every level above the 10th percentile releases per master, cities outside the US have significantly more releases than those in the US. Furthermore, I cannot fully rule out the possibility that releases per master vary both spatially and temporally but thusfar I have not identified a plausible source of bias here besides the interactions of the issues mentioned previously.

The labels file contains information on all the companies associated with the production of the album. This includes labels in the conventional sense but also lacquer cutting facilities, copyright holders, and publishing houses. Most importantly for our analysis, this file contains information on the recording locations of albums. This encompasses recording studios, live music venues, and festivals.

Lastly, the artists database contains the names and id numbers for all musicians in the Discogs database, regardless of whether they were the main artist on an album or merely a collaborator or session musician on another artists release. A key advantage of Discogs is that the database links the names listed in the liner notes to unique identifiers, eliminating the need for messy data cleaning and merging to construct networks. Discogs also maintains a list of aliases, making it easier for users to ensure that the artist linkage is done correct. Tattersal [Tat17] suggests that Discogs artist harmonization is even superior to the scientific author id system ORCiD, which is used extensively in metascience research. An example of the data we extract, as seen in the Discogs UI, is in appendix A.1.

3.2 Geocoding

We assign locations to projects using the following procedure. 6% of releases have a "Recorded At" tag in Discogs. That tag links to a unique identifier in the Labels database. From here we use two sources of information to geocode. In the majority of cases, the "contact information" contains an address or addresses of the studio. After basic cleaning the contact information to isolate the address, we pass the

Geocoding Method	Count
Not Identified	10028
Precise address geocoded with OSM	13457
City parsed from profile	1723
Imprecise address searched using Google Maps API	4709

Table 1: Geocoding Methods used to identify studio locations

text into the Open Street Maps Nominatim Geocoder which returns longitude and latitude coordinates of the studio. This locates 45% of our data. A secondary method relies on the "profile" tag in the data. Frequently, this profile includes location information. Simply parsing the text for geographic names immediately following the phrase "located in" (or semantically similar phrases) boosts location coverage further. Table 1 shows the breakdown of the techniques. We aggregate from albums to releases simply taking any available studio listed in any release unless the "Complete and Correct" data quality indicator is present on that particular release, as Discogs does not have a way to discern the quality of location data otherwise. Where these methods fail, I use the Google Maps API as a more flexible search option for locating studios.

To attach a city to each release, we join the coordinates from the geolocator to the morphological urban areas as provided by Kelso et. al [KP12]. This dataset also contains population information for the precise area in the year 2010, which is carried over into the analysis. These boundaries are drawn using morphological areas from satellite data derived from LandScan, a high resolution remote sensing dataset. Morphological areas used because criteria to produce the boundaries can be universalised across countries rather than reliant on particular administrative boundaries. Studios that do not map to an urban area are not included in the final city-level database.

Approximately 9% of albums in the data record in multiple cities. In this case, the albums is counted in full in both cities. In contrast to Watson (2012) [Wat12], I find that the vast majority of intercity linkages are in geographically neighboring cities. This discrepancy these two findings is due to both the larger sample timeframe of this paper and the focus beyond just charting albums, which are disproportionately likely to have recording locations far apart.

One initial concern would be studios moving. In theory, our data could more precisely account for this, as Discogs documents location changes and notes the years in which the studios were active in particular locations. However, close inspection of studios with multiple addresses finds no instances of studios moving across metropolitan areas so the reconciliation of multiple addresses is not conducted in this study. Instead, studios under the same ownership shifting between cities is simply recorded as separate studios. This does matter slightly in the data however, as studios from the labels data may produce multiple sets of coordinates. In this case, we can simply drop duplicate 'studio_id' city pairs.

3.3 Coverage and Quality

Despite its crowd-sourced nature, the accuracy and completeness of Discogs is considered to be very high [BS17] [Tat17]. Discogs has a robust moderation system including tags for "Correct" and "Correct and Complete" to verify release information. The large number of users also ensures accuracy as audiophiles can comment or edit where information appears inaccurate[2]. Data mining from wiki-like archives similar to Discogs has precedent in the literature [SMT+19]. However, it is important to note that much less is known about the accuracy and completeness of the label information as it is not a core function of the database.

We can understand coverage gaps at three levels. First, a very small fraction of all the music ever made is recorded. The location of recording is not always that location we want to associate with a particular music scene. The development of Afrobeat [3], an offshoot of funk, is credited to the cities of Accra and Lagos, but the vast majority of records were produced in London, LA, and New York because of the quality of studios available compared to those in Africa. Here the data reflects the

[2]There is no way to verify the extent to which this wisdom of the crowds holds for Discogs. However, one piece of anecdotal evidence is particularly striking: an album in the author's collection is limited release with only 500 copies of the album sold. Yet, 216 Discogs users report owning this album.

[3]Discogs provides explainers for each style listed in the database. This is the main source used for style or genre-specific information unless another source is cited.

reality of recording locations, but there are conceptual questions on whether recording studio locations accurately describe the geography associated with this genre.

Second, many albums do not list a recording studio. Conceptually there are two explanations for this. In many cases, attaching a location to an album may not make sense. The low coverage of electronic music is likely due to the fact that home recording is the norm in this genre so describing an album as being made in a particular place is not sensible. Similarly, a large amount of rap music is made asynchronously with producers working separately, though not always independently, from lyricists. More concerningly, albums that ought to be well described with location information may simply not report the studio in the liner notes on Discogs.

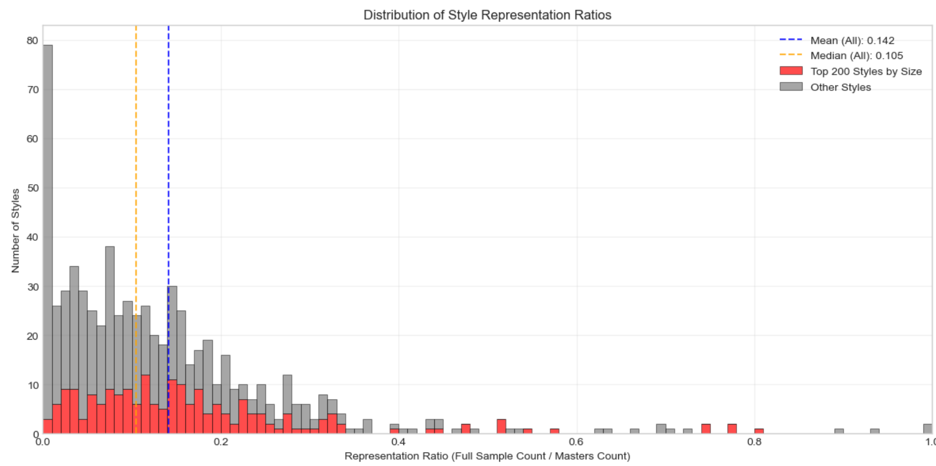

Figure 1: Coverage Rates by 'Style' Tag

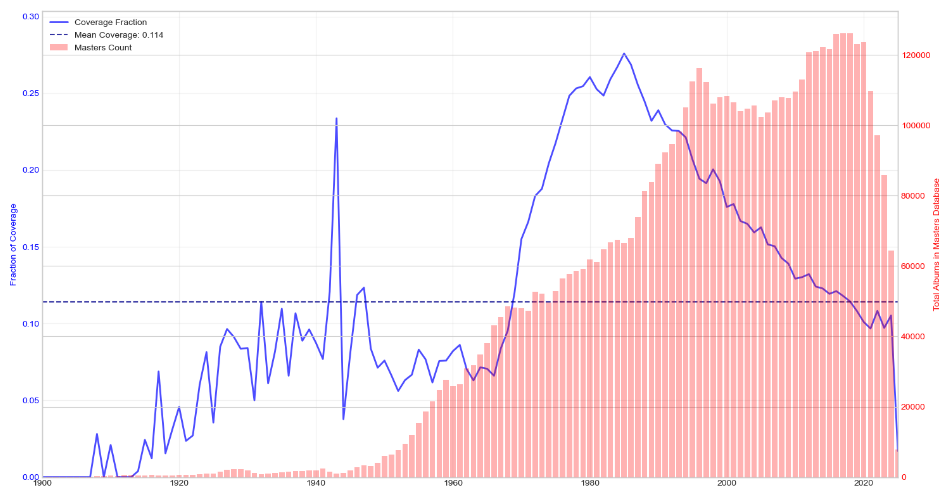

Figure 2: Coverage Rates Over Time

Lastly, we have coverage rates that can be well described by our data. The "recorded at" tag is available for approximately 6% of all releases and approximately 20% of all masters. From our starting point in the masters database we achieve a coverage of 11.4% of all albums. Coverage over time is unexpectedly but unfortunately heterogenous. Figure 2 shows the coverage of the geocoded data compared to the full masters database.

Coverage by genre is listed in Appendix A.2. Unsurprisingly, electronic music is the most poorly represented with only 6% of masters being mapped to an urban area. Classical, jazz, and Rock command much higher coverage at over 15% of all albums being attached to a location. Figure 1 shows the variation in coverage across Discogs. Here, the coverage - the fraction of albums in the full albums database that are matched to an urban area in our cities panel - is shown on the x axis and the frequency is shown on the y axis. At the far left of the graph, some styles have no representation in the data, while a few styles have very high coverage. The largest 200 styles in the histogram are noted in red, suggesting that few larger styles have very low coverage rates.

3.4 City-Level Measurements

To examine the extent to which diversity is a driver of creativity, I construct a genre level Herfindahl-Hirschman Index, defined below. The higher the index score, the more specialized a city is in a particular genre.

$$\text{HHI} = \sum_{i=1}^{N} g_i^2$$

Where g is the share of a given genre from N genres.

To enrich the basic statistics, I construct a network of the cities as nodes in a graph. I leverage the dataset's ability to track artist locations over time to build a directed, weighted network of cities using the sum of artist 'transitions' from city to city. In this case, a transition is merely the presence of a musician in city B after city A, which corresponds to an edge from A to B with weight 1. We compute betweenness centrality using the Python package NetworkX [HSS08]. Betweenness centrality is described by:

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t \mid v)}{\sigma(s,t)}$$

where v is the city-node from the set of V city-nodes. $\sigma(s,t \mid v)$ describes the number of shortest paths in the graph passing through node v as a fraction of the total shortest paths described by $\sigma(s,t)$.

This setup is different to Watson (2012)'s city network [Wat12] which utilises an undirected graph weighted by the sum of albums with recording in both cities; a city pair gets a weight 1 if a given album lists two studios, one in each city. The interpretations of centrality between this paper and Watson are only somewhat different. The betweenness centrality measured in this paper identifies cities that facilitate connections between artists originating in different cities. Watson's centrality, meanwhile, identifies cities that facilitate connections within individual projects.

3.5 Intra-City Graph Measurements

To build the network of musicians for a given city, we begin by parsing the releases data for the "extraartists" tag which is used to indicate the musicians involved on a given project and their role. Common role tags include "Bass", "Mastered By", "Art Direction", "Conductor", etc. The artists these tags refer to link to a unique id, bypassing the need for tricky and unreliable cleaning to distinguish between multiple artists of the same name.

Because we want to reflect the network of musicians present in a given city at a particular point, we filter out edges from the graph that are unlikely to reflect direct collaboration. On an unfiltered graph lacquer engineers, technicians who prepare vinyl for pressing, have extremely high node centralities. This is mostly due to the high throughput of production facilities relative to recording facilities leading the node degree of lacquer cutters to be extremely high relative to individuals in the creative process. Since this and similar roles such as legal teams are not directly related to the creative process, these are filtered out.

Secondly, we filter out songwriting and lyrics credits. While many of these credits reflect real collaboration, these tags are very often used to give writing credits on covers. This step is taken in contrast to Budner and Grahl [BG16] who explicitly include writing credits [4]. However, because this dataset does not filter albums to a few influential releases, cover versions overwhelm the data. The result is that unfiltered networks indicate very high centrality for musicians located outside of a given time period, particularly in genres that rely heavily on a standard songbook like Blues and Jazz.

The resulting data is a large list of artist linkages to release ids. I remove duplicates, which are particularly common as some releases list the personnel on every individual song rather than the

[4]Gienapp et al, the other paper using discogs to construct collaboration networks, make no mention of cleaning writing credits.

album as a whole. This may result in over-emphasizing tangential collaborators relative to more core participants. However, since the role tag leaves us no method of determining the relative importance of a given role, it would be ill-advised to attempt to assign importance. To construct the within-city networks, the release-artist links are filtered by city and time period and then joined many to many to produce artist-artist links which are counted. Every collaborator on an album is matched with every other collaborator on the same album, in keeping with the approach used by Gienapp et al [GKB21]. The result is an undirected, weighted graph where the weights reflect the number of albums where the pair of artists collaborated.

I generate a graph for each city in decade chunks beginning with 1950 and compute the statistics for each period. For robustness, I also compute the "off-decade" statstics for 10 year spans beginning in 1955. In addition to counting nodes and edges, I compute the graph density. Graph density is defined as the fraction of possible node to node pairs that are realized in the network. It is a measurement of the extent of collaboration in a given city. A demonstration of network density can be seen in Figure 3.

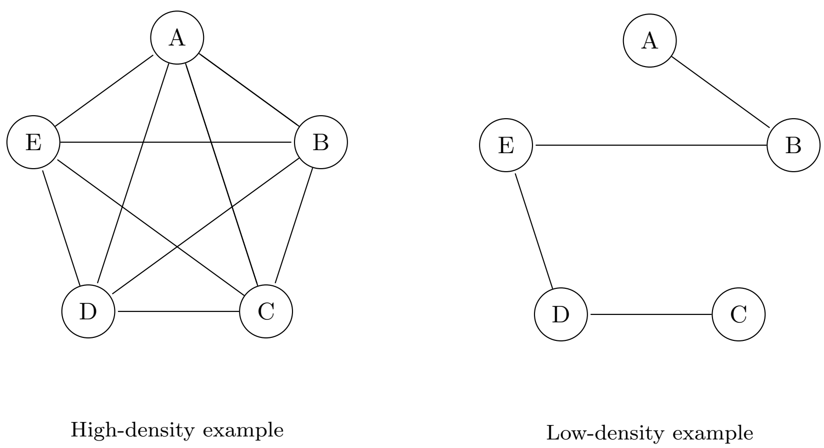

Figure 3: Examples of high-density and low-density undirected networks.

In both cases, our regressions on network variables always control for size, either that of the number of albums from the city or the number of collaborators - nodes - present in the data. This is because both measurements are highly sensitive to network size. Two important caveats regarding sampling are needed. Because the graphs for each city are constructed using the sample of albums where the city is known. The graph is inherently missing a large number of nodes and edges. Two nodes within a city may be connected by nodes that operate outside of that city - this is not captured. As with other points in the data, we are also missing nodes and edges from albums that are note geocoded. Generally speaking, the coverage of the personnel in Discogs is vastly superior to geographic coverage.

3.6 Measuring Innovation

To measure musical innovation we borrow from patent literature. Kelly et. al [KPST21] construct an index of innovation at the patent level using a natural language processing approach. First, they identify important words in a patent document using a term-frequency inverse document frequency approach, though modified slightly such that the inverse document frequency only considers prior patents. From there, they take this vector of words in the patent are compute the cosine pairwise similarity between this patent vector and the vectors of all previous and then all future patents. Their resulting measurement is the ratio of the average similarity to future patents divided by the average similarity to prior patents.

The intuition is that innovating patents are those that are very different from patents that came before it and very similar to patents that came after it. This algorithm has been used outside of patent literature, notably in Voth and Yanagizawa-Drott [YDV24], who deploy the algorithm in pursuit of measuring cultural innovation in fashion. Their unit of analysis is a very large panel of high school yearbook photos, with each photo turned into a vector of fashion characteristics - hair length, shirt pattern, etc.

The innovation algorithm in this paper uses the 'styles' as the input vector for comparisons. Innovative albums are those that feature tags or combinations of tags that are both novel and go on to become popular. The computations are run at the masters level and aggregated up to the city-decade level. Mechanically, the algorithm behaves more like Kelly et. al (2021) than Voth and Yanagizawa-Drott (2024); the larger number of tags means that the measurement has much more right skew than in the latter paper. Since there is no way to determine the relative importance of tags within an album, I do not use the TFBIDF normalization prior to the similarity calculations as done by Kelly et al. The advantage of this approach is that I can fluidly capture both the introduction of new musical styles and the introduction of novel combinations of musical styles. A score of 1 indicates that a given album is as similar to previous records as it is to future ones. Since new styles get added over time, these innovation scores average more than 1 in the majority of cases. Additionally, the possible scaling from zero to infinity means that the distribution is strongly right-skewed. While it would be possible to normalize this, I opt to continue the strongly skewed distribution to better discern where the highest innovation lies.

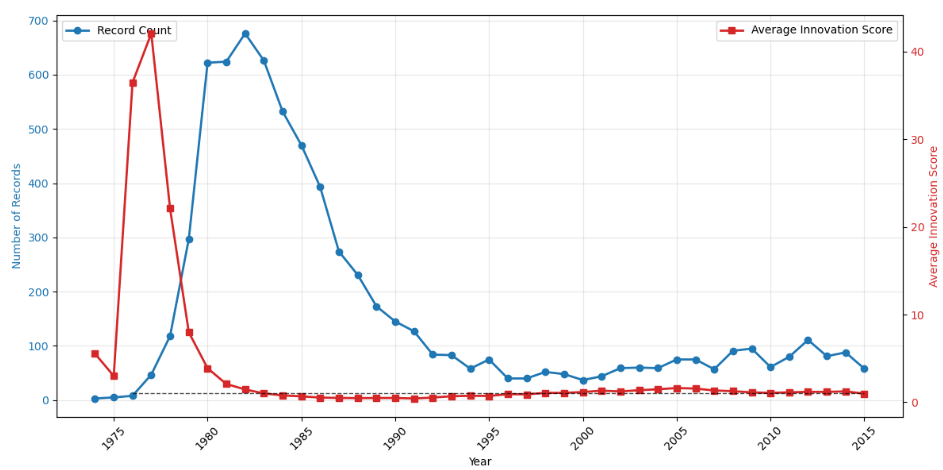

Figure 4: Volume and Average Innovation Scores for "New Wave"

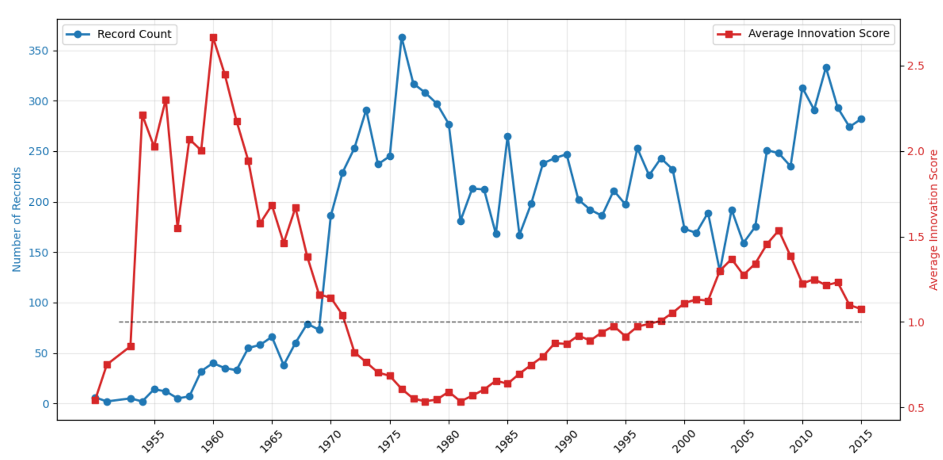

Figure 5: Volume and Average Innovation Scores for "Folk"

Figure 4 depicts how our innovation measurement scores albums with the "New Wave" tag across time, compared to the volume of releases. This particular tag follows a trend similar to many in our data. A very sharp uptake in the style followed by a decline as tastes change. Our innovation score peaks immediately prior to the peak in the volume of releases. Late 70s New Wave in particular produces some of the highest innovation scores in the data on account of the large volume of early 80s New Wave in the data. Among the 5 highest scoring major albums (defined by 10 or more releases), 3 are New Wave records from this era [5]. It is important to note in these graphs that the innovation score depicted in each graph is not the score derived mechanically from each tag but rather the average score for albums with that tag; the other genres attached to an album will influence the score here.

Figure 5 is constructed similarly to figure 4. This tag, however, has a much more complex trajectory over time. Our innovation measure still indicates high values prior to upswings in the popularity of that genre. Its clear from this particular graph that "innovation" in this sense is not precisely innovation but a measure of whether an album is a forerunner of trends. For example, it's not clear from this measure alone whether the "highly innovative" folk musicians of the mid-2000s are actually bringing new ideas to the genre or whether they happen to be making the right type of music prior to a trend they did not influence.

Important caveats to this innovation score need to be made. Firstly, this index only captures a small aspect of innovation. Aspects of the music that cannot be represented by these tags, such as innovations in recording techniques, lyricism, or instrumentation, are not captured. Even within the strict definition of style leaders and laggards, individual albums are incredibly sensitive to the highly subjective labels placed on them by the listeners of that album. Individual innovation scores are not particularly meaningful until they are aggregated across cities. The nature of these style tags is also inherently backward looking. Genres and subgenres are generally defined in hindsight as people look to define particular trends. Two pieces of evidence in this data support this hypothesis. First 7 shows us that innovation is muted in the years since 2000. Second, ordering the albums by innovation score indicates that the first albums or singles by an artist tend to be scored as more innovative than their later output.

3.7 Empirical Strategy

To test for determinants of innovation, we use a set of fixed-effects regressions on the city-decade panel. While the innovation scores is generated at the album level, the independent variables we test are all at the city level. In the main specifications, I lag the independent variables of interest in order to avoid contemporaneous effects that may bias the results [BL]. The baseline specification is described:

$$y_{ct} = \beta x_{c,t-1} + \gamma_c + \delta_t + G_T + \varepsilon_{ct} \tag{1}$$

where y_{ct} is the innovation outcome for city c in decade t, $X_{c,t-1}$ is the lagged independent variables, γ_c are city fixed effects, δ_t are decade fixed effects, G is a vector of genre shares in time t and ε_{ct} is the error term.

Time fixed effects capture both changes in the overall level of innovation as well as time variation that could be caused by changes in the composition of the underlying Discogs database. For example, we may suspect that changes in predominate listening formats affect the number of releases, as well as the scope of information contained in their liner notes. City fixed effects soak up effects caused by stable characteristics in the city. Stable country-by-country differences in musical production that do not vary over time such are captured here.

I test the average innovation score, as well as innovation scores at the 75th and 95th percentile. Prior qualitative research suggest that older traditional styles and modern cutting edge sounds are not substitutes but compliments [**?**]. The percentile approach allows us to distinguish between general effects on the level of musical innovation compared to the effect on the most innovative albums.

For consistency, our main specifications are restricted to any city over 200 total albums across time periods. Some robustness tests at different cutoffs are conducted where the data allows it. The rationale for this cutoff is to ensure that we don't attempt to generate and analyse intra-city graph measurements for networks that are too small. While the cutoff is arbitrary, this number was chosen to roughly balance the known baselines for music collaboration networks described in the literature, 500 albums in the case of Budner and Grahl [BG16] and the roughly 50 albums per network from Watson [Wat12].

[5] Those albums are Elvis Costello's My Aim is True, Blondie's Self-Titled, and Patti Smith's Live at the Roundhouse London 1976 at 1,4, and 5 respectively. Individual album's innovation scores should be interpreted cautiously, however, both for the limitations of the style tags as previously discussed and for the fact that forward and backward similarity is taken against a random sample of past and future releases.

Additionally, I control for the contemporaneous genre share as different genres vary categorically in their innovation scores. This may be due to a larger number of styles that can describe the genre or uncontrolled OLS on innovation may capture changes in genre composition rather than the intended effect. Since styles are not strict subsets of genres, this is unlikely to be co-linear with innovation scores. Lastly, I cluster standard errors at the city level to account for potential heteroskedasticity across cities as well as within-city serial correlation in the error term. Coverage gaps may be more pronounced in smaller cities, which can lead to greater variance in innovation measures. While the forward- and backward-looking nature of the innovation score draws from releases across all cities, persistence in local music scenes means measures may still be correlated over time within a city. Rerunning the regressions with and without clustering yields similar coefficients and standard errors, suggesting that clustering does not substantially alter the results but provides a conservative correction.

4 Results

4.1 Innovation descriptives

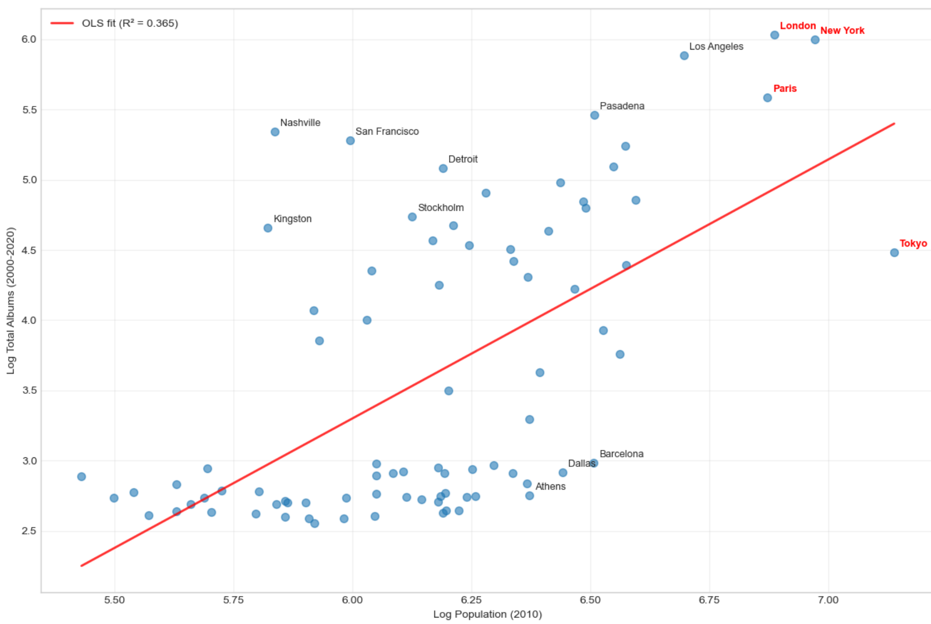

Figure 6: Relationship between Population and Total Musical Output. Largest outliers are listed in black while the four largest cities by population are listed in red.

Figure 7 shows the innovation index scores over time across all albums. The solid blue line depicts the average innovation scores alongside the 75th, 90th, and 95th percentile scores. There is substantial heterogeneity across time with spikes coinciding with many periods widely understood to be innovative periods in music. The spike in the mid-50s coincides with the entrance of Rock n Roll into the mainstream. Mirroring the findings of Voth and Yanagizawa-Drott [YDV24], I find the late 60s to be a particularly innovative period. Spikes in recent years appear more muted. This may be due to an increase in the number of style descriptors available. Unfortuantely this could also be due to sampling - coverage of older releases may be more limited to "important" albums that are better cataloged and preserved. This graph also suggests some degree of cyclicality. Figure 8 depicts the decomposition of these innovation scores into their forward and backward components. The effects of database coverage are clearly visible with both backward and forward similarity being much higher in the early data compared to later data.

Moving to a geographic view, London, New York, and Los Angeles are the clear frontrunners in the volume of music production. Paris is a distant 4th with Nashville and San Francisco further behind. Figure 6 shows the relationship between population in 2010 in the precise urban area as defined by our urban boundaries from Kelso [KP12] and musical output in the period from 2000 to 2020.

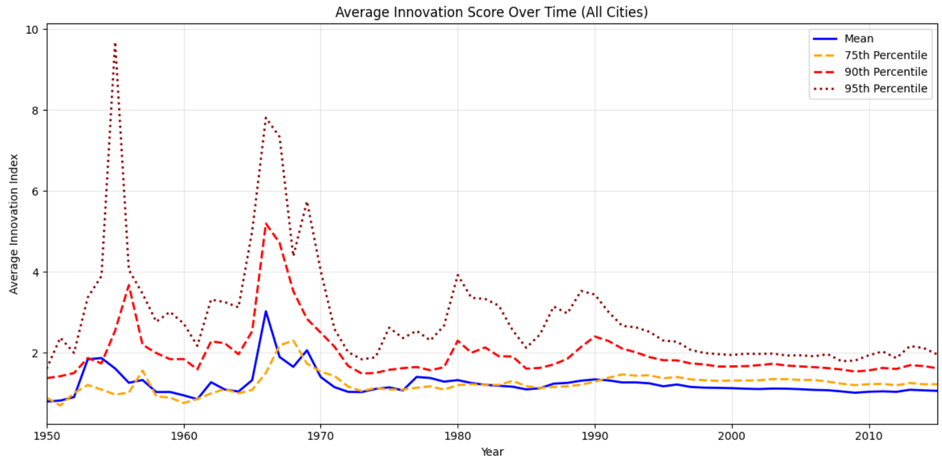

Figure 7: Average Innovation Over Time, Aggregated from Album Scores.

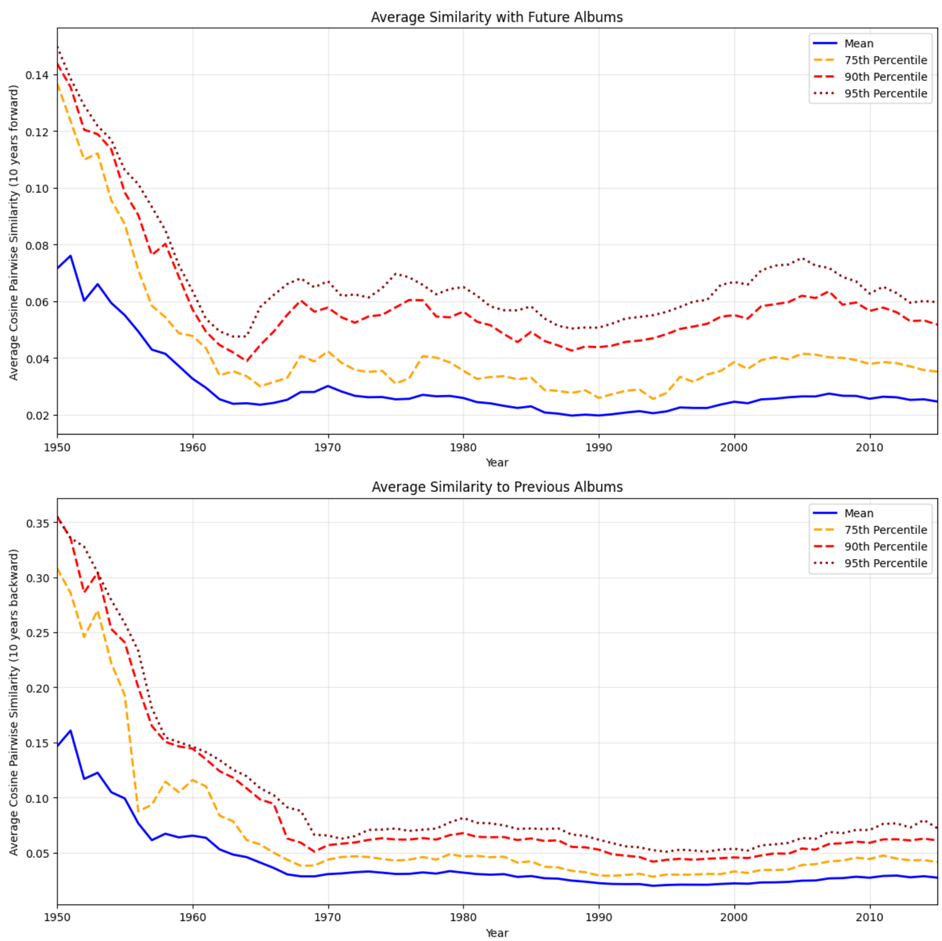

Figure 8: Decomposition of Innovation Scores Over Time

Known music powerhouses like Kingston, San Francisco, and Detroit join Nashville in being highly productive relative to their population size. Meanwhile several non-anglophone cities such as Athens, Tokyo, and Barcelona are represented poorly relative to their size. Figure 9 shows the relationship between the size and diversity of genre output. Kingston, Jamaica is the clear outlier here being well specialised in Reggae and its many subgenres.

Table 2 shows the top 10 cities by average innovation aggregated across time. Tampa's position at

Figure 9: Genre Diversity Compared to Musical Output

the top may be suprising to most audiophiles but not to metal-heads; in the late 80s and 90s Tampa gained a reputation as the "death metal capital of the world" (Swiniartzki, 2021 [Swi]). It stands out both on average innovation levels and specialisation. Its representation in the dataset is dominated by Death Metal and other subgenres of extreme metal. I also find particularly strong innovation scores in the North of England. This stands in stark contrast to the Nesta report [BDFH15] which suggests a very high degree of concentration in music industry employment London and the southeast. Figure 10 suggests that this finding is not driven by particular genres, but rather reflects innovation across a diverse range of music.

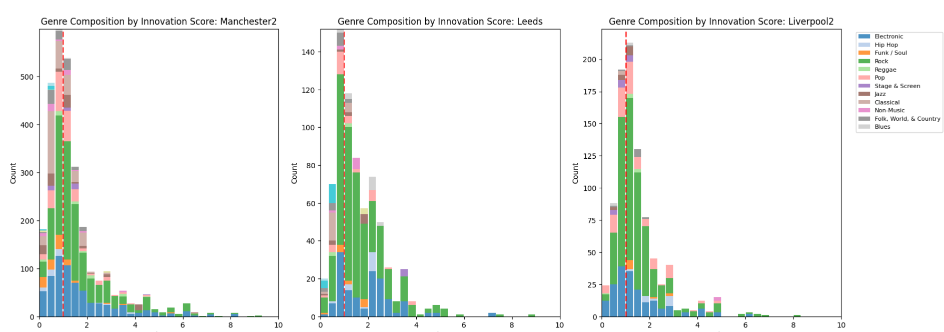

Figure 10: Genre Composition of Select Northern Cities, Arranged by Innovation Score

Evidence for the "Nursery Cities" hypothesis is suggested by examining the evolution of styles over time. We find that as a given style matures, the geographic dispersion of that style increases. The figure 11 depicts an the average trend in concentration, measured by the Herfindahl-Hirschman Index, across all styles in our database. As styles evolve, we observe falling geographic concentration consistent with diffusion of musical ideas.

City	Score	Styles Associated with Local High Innovation
Tampa	2.392	Death Metal, Grindcore
Newcastle	1.847	Heavy Metal, Black Metal
Leeds	1.842	Death Metal, Synthpop
Auckland	1.749	Post-Rock, Indie Rock
Manchester	1.741	Indie Rock, New Wave
Evanston	1.686	House, Acid House
Liverpool	1.685	Indie Rock, New Wave
Düsseldorf	1.678	Euro-House, House
Alexandria	1.675	Punk, New Wave
Chicago	1.669	House, Deep House

Table 2: Top 10 Cities by Innovation Index

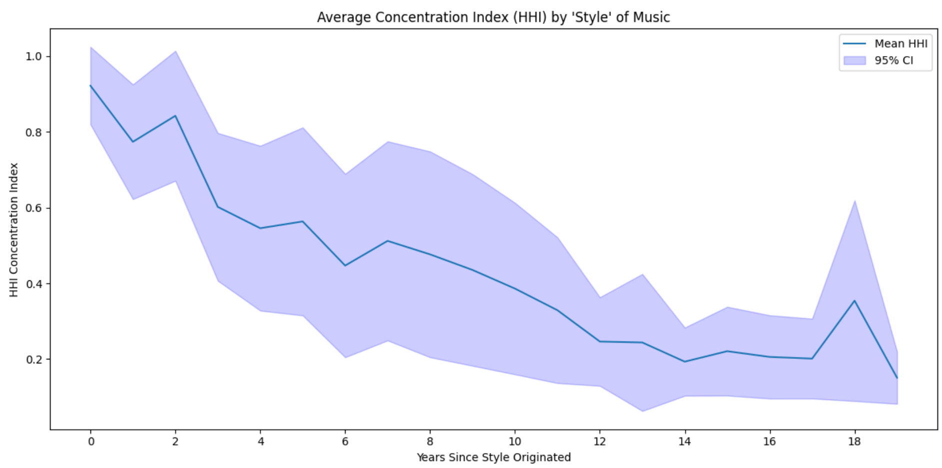

Figure 11: Evolution of geographic concentration over time, aggregates of all musical 'styles'

4.2 Innovation and Agglomeration

I find that musical innovation in a city is significantly negatively correlated with the size of that city's musical output. This finding holds for both the average level of innovation and the 75th, 90th, and 95th percentiles of innovation, though there is heterogeneity. Elasticities of average innovation scores with respect to size is -.129. A 1% increase in total output is associated with an 0.083% decline in innovation scores at the 75th percentile and a 0.203% decline at the 95th percentile. Results are shown in Table 3.

One might suspect a compositional effect - artists from small cities are more likely to release their first album there before moving to a large city. This is suggested by the positioning of large cities in our centrality network. First releases are generally scored as more innovative than subsequent releases. This could be due to artists being genuinely more creative in their early output. Equally likely is that labels are applied retroactively to artist's early output where it may not otherwise fit; artists who espouse a given style later in their career may induce listeners to attach labels that would appear out of place on other artists in that early period, making our innovation measure appear higher. To test this, I create an average in each city period from only those releases which coincide with the first album from a given artist. We produce an elasticity of -0.09, not dissimilar from the baseline innovation results.

For additional robustness we test the "off-decade" setup with periods shifted 5 years forward from the main setup. The results are slightly smaller but remain highly negative and significant. This finding is also robust to changes in the size cutoffs. Expanding our range to all cities, raising the cutoff from 200 to 500, and dropping observations above 1000 albums per decade all fail to negate the main finding.

As a further robustness test, we exclude 10 satellite cities from our dataset that could plausibly be considered part of a consolidated labour market with their larger city. Artists in large cities often record

Table 3: Effect of Total Musical Output on Innovation

	Avg. Innovation Score of First Releases (log)	Avg. Innovation Score (log)	Innovation Score 75th percentile (log)	Innovation Score 95th percentile (log)
Lagged log # Albums	-0.0943***	-0.129***	-0.0834***	-0.203***
	(0.0307)	(0.0185)	(0.0202)	(0.0261)
Constant	5.146	-0.834	-1.428	-2.027
	(3.341)	(1.395)	(1.393)	(2.719)
Genre Controls	Yes	Yes	Yes	Yes
Observations	348	425	425	425
R-squared	0.169	0.326	0.202	0.320

Standard errors in parentheses
Fixed effects included
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

at studios further from the city where studio time may be comparatively less expensive; its possible that this is especially true for albums with more intensive recording processes which may in turn result in more innovative albums. This is especially problematic given the city shapefiles available. The city definitions I use from Kelso et al [KP12] are morphologic but still attempt to distinguish adjoining satellite cities, resulting in imprecise boundary delineations. The full list of excluded cities is in Appendix A.5. Yet, eliminating these adjoining cities from our data doesn't weaken the correlation. The precise results from this regression can be found in appendix A.6

Furthermore, I divide the data into US and non-US cities. Within the 38 American cities in our dataset, we see an even stronger negative effects of total city output on innovation. The effect is much weaker for non-US cities with correlations breaking down at the 75th percentile and for the average innovation scores of first releases.

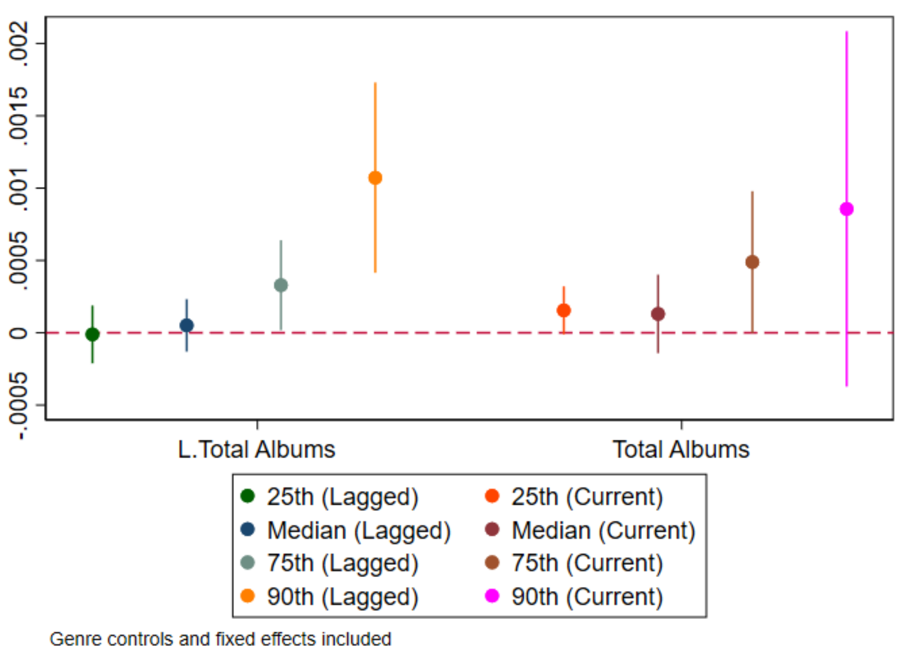

Figure 12: Coefficients of Lagged and Contemporaneous Music Output on Commerical Success

However, while innovation seems to be optimised in small cities, the opposite is true for the commercial success. Using the number of releases per master as a proxy for market distribution, we find that commercial success is strongly correlate with lagged city output after controlling for contemporaneous genre distributions. However, it appears this effect is driven by the top end of the distribution -

the 70th and the 90th percentile number of releases per master are significant but no lower percentiles are. Figure 12 shows the coefficients in the baseline specification. However, this finding is not robust to the exclusion of satellite cities as described previously.

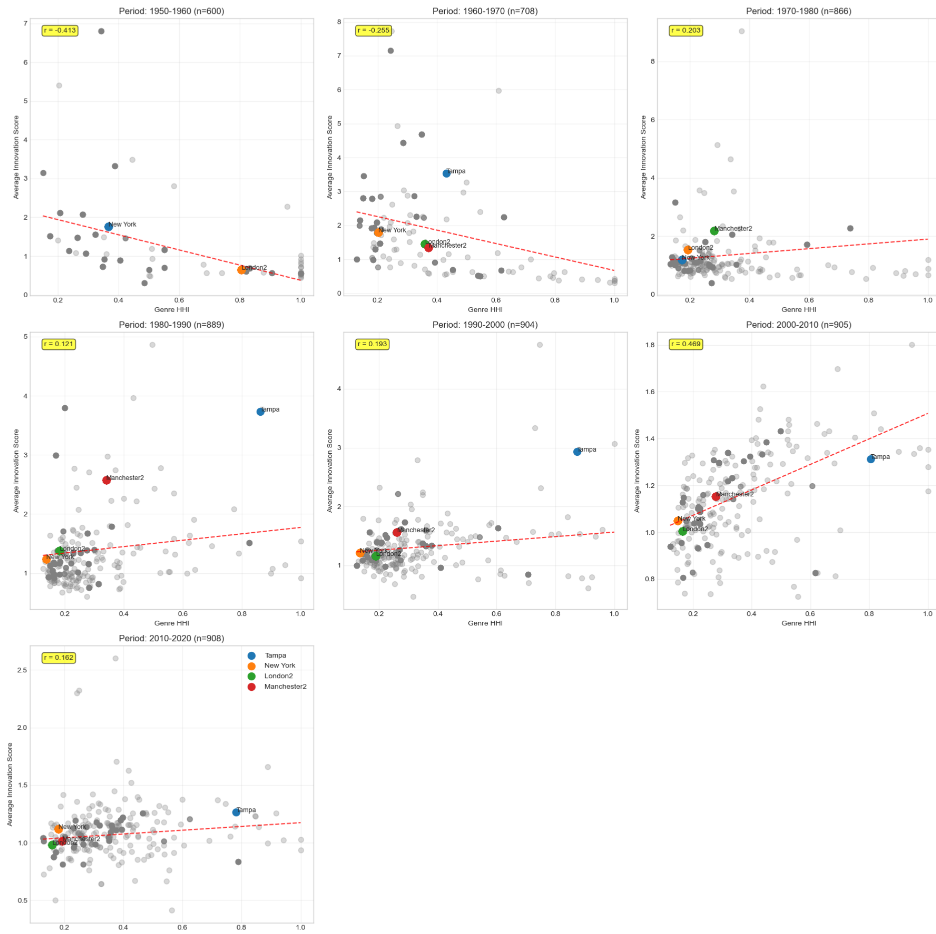

Figure 13: Decade-by-Decade shifts in the relationship between musical diversity and the average innovation score. The two largest markets (New York and London) are identified alongside two particularly innovative cities (Tampa and Manchester). A downward slope suggests that diverse cities are more innovative while and upward slope suggests that specialized cities are more innovative.

I find no statistically significant effect of diversity on innovation, where diversity is described by the Herfindahl-Hirschman index of genre shares within a city. This does exhibit some interesting time heterogeneity. Figure 13 depicts the relationship in each decade between the innovation and the diversity of musical output. Unlike in our regression analysis, Figure 13 depicts contemporaneous, rather than lagged, diversity. While not a robust statistical test, the graph generally suggests a changing relationship between specialization and innovation.

4.3 Network Effects

Figure 14 shows the changes in the ordinal rankings of centrality of artist movements. By and large the central cities in the network are the same urban areas that command large musical output. There are some exceptions: Tokyo appears more central in the network than its output size would suggest. The big three cities - London, New York, and Los Angeles - monopolize the top spots across the all periods in the data. Early decades suggest higher centrality of European cities

I find that centrality is negatively with lagged centrality. This suggests that high inflows of musicians does not induce greater creativity. The negative correlation is most likely to the compositional effect; we do not observe that first releases are significantly less innovative. Dividing the innovation

Table 4: Effect of City Centrality on Innovation and its Component Similarities

	Avg. Innovation Score (log)	Avg. Innovation Score of First Releases (log)	Avg. Innovation Score (log)	Avg. Forward Similarity (log)	Avg. Backward Similarity (log)
L.Centrality	-0.191	-0.749	-3.724**	2.128**	5.184***
	(1.906)	(2.571)	(1.714)	(0.884)	(1.593)
L.Total Albums		-0.000356***	-0.000385***	-0.0000156	0.0000664
		(0.0000889)	(0.0000892)	(0.0000288)	(0.0000538)
Constant	-0.590	3.778	-0.728	-4.412***	-4.894***
	(1.379)	(2.759)	(1.202)	(1.117)	(1.567)
Genre Controls	Yes	Yes	Yes	Yes	Yes
Observations	440	358	440	440	440
R-squared	0.131	0.155	0.203	0.494	0.404

Standard errors in parentheses

Fixed effects included

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Effect of Density on Innovation

	Average Innovation Scores	Average Innovation Scores	Average Innovation Scores	Average Innovation Scores
Network Density	0.613**			
	(0.291)			
L.Network Density		1.220***	1.223***	1.229***
		(0.388)	(0.387)	(0.390)
L.Total Albums		-0.000420***	-0.000184	
		(0.000106)	(0.000144)	
Total Albums				0.0000778
				(0.000204)
Number of Nodes			-0.0000217**	-0.0000379***
			(0.00000987)	(0.00000872)
Constant	1.249***	0.830	0.808	0.786
	(0.0366)	(2.762)	(2.772)	(2.764)
Genre Controls	No	Yes	Yes	Yes
Observations	508	472	472	472
R-squared	0.0213	0.247	0.250	0.249

Standard errors in parentheses

Fixed effects included

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 14: Trends in the Ordinal Rankings of Centrality as Measured by Artist Movements

score into its component parts we see that centrality appears highly correlated with both the average forward similarity and average backward similarity. This is shown in table 14. I interpret this as cities more central as being more "on-trend" but not more ahead of trends.

We find that innovation is significantly positively correlated with network density in given cities. This holds when we control for contemporaneous genre composition and lagged city total output. The results are shown in table 5. The effect of network density on innovation holds for almost all the robustness tests performed in the prior assessment of size [6]. The findings are robust to the exclusion of satellite cities are strengthened substantially by isolating cities in the US. For non-US cities, only the coefficient of density on the 95th percentile innovation score holds. When we isolate the smaller end of our data by eliminating cities over 1000 total albums, the effect is strengthened. However, raising the size cutoff from 200 to 500 eliminates the effect.

5 Discussion

These findings largely contradict the conventional view of creative geography that large, diverse labour pools nurture creativity. The most 'creative' cities in this data are not significantly more musically diverse. Innovative hubs identified in the data are a different set of cities from those identified by previous research as hubs for the industry [FJ10] [Wat12]. Measurements of centrality within a global network of artist movements largely confirms these assessments, with traditional

The finding that smaller cities appear more innovative has a number of possible explanations. The first and most plausible is that knowledge diffuses much more freely in music than in other forms of creativity. Unlike commercial innovation, musicians seek to distribute their ideas as widely as possible. Touring and album sales rapidly disseminate musical ideas not long after acts debut on the local circuit. This opens up further research possibilities on empirically studying the dissemination of music as it pertains to musical output. Discogs itself records the country of particular releases, allowing for a coarse measure of dissemination internationally. For a finer geographic resolution, databases of touring schedules such as Setlist.fm or Spotify's publicly available touring data [Li] could be tied to the musical output data from discogs to see if there is a concrete link between the touring destinations of genre forerunners and subsequent albums in that genre.

A large number of artists playing in a given area and a given genre may enforce orthodoxy rather than allow artistic liberty. This hypothesis needs further testing but we find some evidence in our analysis of centrality in the global network of music. Central points in the network produce music

[6] We cannot run a robustness test where we drop the below the 200 album per decade threshold, as we do not generate graph statistics for these cities.

that is both similar to past trends and indicative of future trends. The centrality results from table 14 provide limited evidence for this, suggesting that overall cities core to the global music network are more on-trend than small cities, but less ahead of trends. While competition is generally though to be a driver of innovation in conventional industries, artistic creativity may not benefit from competition in the same sense.

Another consideration is that nature of labels. The geographic concentration of musical employment identified in prior literature [BHM] is largely a function of concentration of more traditional commercial components of the music industry. Even casual music listeners are likely familiar with the tension between the artistic and commercial aspects of music making. Large labels are often seen as risk-adverse and unwilling to permit artistic experimentation [Aze01]. Stylistic innovation is a very particular aspect of musical innovation; other forms of innovation such as technical skills in recording techniques may be more sensitive to agglomeration.

A second plausible explanation relies on that fact that unlike the technical patent text that underpin Kelly et.al's [KPST21] analysis, the style markers in this dataset are not completely exogenous to location. Because the idea of local music scenes is ingrained in the popular consciousness, many of the delineations we make regarding genre boundaries are built from identifying clusters of like-minded musicians geographically. This could explain why dense networks are highly correlated with more innovation - communities of musicians could be grouped together as a genre even where they are sonically distinct. For example, "Midwest Emo", "Italo-House", and "Memphis Rap" all appear in the top 30 thirty styles when ranked.

Ranking	Style	Avg. Innovation Score	First Year	Most Common City	Second Common City	Albums in our Sample
3	Midwest Emo	6.165	1995	Chicago	Urbana	6
9	Britcore	4.185	1992	London	Berlin	6
20	Balearic	3.039	1986	Paris	London	20
24	Italo-House	2.679	1988	Milan	Rome	206
25	Memphis Rap	2.650	1991	Memphis	Houston	9

Table 6: Style Tags with a Plausible Geographic Origin. The rankings on the far left indicate the ordinal position in the full list of styles arranged by average innovation score.

Mechanically, its not clear the extent to which endogenous labeling is an issue in calculating an innovation score. A style label applied to a single local cluster will not produce a high innovation score if that genre stays relatively insular - forward similarity doesn't ever get particularly high. If that regional style does diffuse more broadly, than the innovation score is correctly capturing that cluster as a fore-runner of trends. Ultimately, styles based on geographic characteristics are unlikely to make a significant difference in the aggregate because their size relative to the universe of data is small. This can be seen in Table 5.

Despite the inverse size finding, local density measurements appear to be a driver of innovation. The localized positive density response, coupled with the negative centrality response in the intercity network, supports the scale-dependent density effect found Guler and Nekar 2010 [GN12]. Local face-to-face collaborations seem to be a good for the creative process, while centrality effects at a large scale fail to induce creativity.

6 Future Research

By far the area in need of the most futher study is the extent to which the 11% of albums successfully geocoded are representative of the reality of musical geography. Whether an album chooses to note the recording studio is not a random choice. Increasing the coverage of the data would go the furthest toward overcoming key limitations. Our specifications are city-decade and rely on lagged independent variables, meaning that the characteristics of music in 1989 could be correlated with events and output happening in that city as early as 1970. This may make more conceptual sense for independent variables such as musical output and less for graph measurements where bands and musicians have completely come and gone between a given period and the prior period. We can increase the precision of our results by cutting finer temporal resolution, but this necessitates more data.

Further efforts could expand and refine this geolocation. Since coverage of personnel on an album is better than coverage of recording studios, it would be possible to infer the location of the studio based on the presence of particular engineers or personnel. Graph representation techniques could be refined to predict the location given the releases position in the collaboration network. Ultimately this imputation is not undertaken because it would bias the subsequent regressions that utilize properties of the network within cities.

The distribution of highly innovative styles suggests that rock and to a less extent electronic music may be a disproportionate driver of our innovation metric. This is likely a result from coverage. Styles that emerge for the first time during our sample period cause much higher innovation scores. Future research should also test the extent to which innovation is sensitive to the tag set used. One could augment the style tags using those from other datasets such as Spotify or RateYourMusic, both of which include richer set of qualitative descriptors. Beside this, a better understanding about how the Kelly algorithm responds to dropouts to shifts in the set of tags available. A simple test would be re-running the algorithm several times with styles randomly dropped from the dataset and confirming that it produces the same results when aggregated.

Another concern is the city delineations. Further tests for the result's sensitivity to urban boundaries are needed. This is particularly pertinent given issues with satellite city delineates as described in Section 3.2. A good approach to rectify this most precisely would be to isolate particular countries where units such as the US's Mean Statistical Areas (MSAs) can be used with greater accuracy.

The flexibility of the innovation algorithm borrowed from Kelly et. al can be leveraged using a wider variety of artistic data. While it has yet to broach the subject of geography, there is a fledgling literature seeks to mathematically describe relationships between musical pieces in a corpus of songs. Tsai and Ji [TJ20] and Serra et al (2021) [SSC21] take an approach of vectorizing sheet music, allowing music to be described using the types of structures suitable for natural language processing. In the former study, they leverage this approach for classification. The latter paper uses this approach to construct a time-series measurement of harmonic complexity of music; adding an ability to attach location information offers the potential for this approach make sheet music an analogous tool for patent text in innovation literature. large corpuses of lyrics, sheet music, or even the MIDI files themselves can be trawled for the innovativeness of creative works.

In this research I demonstrate the capability of a novel database using geocoded album liner notes to describe music production over space and time. My results show that dense networks of musicians in small urban areas produce the most innovative music. This finding appears very robust. In all, the data suggest a triumph of the local music scene.

References

[ANNN21] V.S̃. Araújo, C. M. Neto, R. G. Nakamura, F. and F. Nakamura, E.˙ Using complex networks to assess collaboration in rap music: A study case of dj khaled. *Digital Humanities Quarterly*, 15(1), 2021. Online; referred to as "Araújo et al. 2017" in related literature.

[ARS25] Nicholas Argyres, Luis A. Rios, and Brian S. Silverman. On the heels of giants: Internal network structure and the race to build on prior innovation. *Strategic Management Journal*, 46(7):1728–1761, jul 2025.

[Aub] Florian Aubke. Creative hot spots: A network analysis of german michelin-starred chefs. 23(1):3–14. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/caim.12044.

[Aze01] Michael Azerrad. *Our Band Could Be Your Life: Scenes from the American Indie Underground, 1981–1991*. Little, Brown, Boston, 2001.

[BDFH15] Hasan Bakhshi, John Davies, Alex Freeman, and Peter Higgs. The geography of creativity in the uk: Creative clusters, creative people and creative networks. Research report, Nesta & Creative England, 2015.

[BG16] Pascal Budner and Jörn Grahl. Collaboration networks in the music industry. *CoRR*, abs/1611.00377, 2016. arXiv preprint.

[BG21] Enrico Berkes and Ruben Gaetani. The geography of unconventional innovation. *The Economic Journal*, 131(636):1466–1514, May 2021.

[BHM] Rafael Boix, José Luis Hervás-Oliver, and Blanca De Miguel-Molina. Micro-geographies of creative industries clusters in europe: From hot spots to assemblages*. 94(4):753–773.

[BL] Stefano Breschi and Camilla Lenzi. Co-invention networks and inventive productivity in US cities. 92:66–75.

[BS17] Dmitry Bogdanov and Xavier Serra. Quantifying music trends and facts using editorial metadata from the discogs database. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017.

[Cas] Carolina Castaldi. The geography of urban innovation beyond patents only: New evidence on large and secondary cities in the united states. 61(7):1248–1272. _eprint: https://doi.org/10.1177/00420980231204718.

[CK] Gerald Carlino and William R. Kerr. Chapter 6 - agglomeration and innovation. In Gilles Duranton, J. Vernon Henderson, and William C. Strange, editors, *Handbook of Regional and Urban Economics*, volume 5 of *Handbook of Regional and Urban Economics*, pages 349–404. Elsevier. ISSN: 1574-0080.

[Coh91] Sara Cohen. *Rock Culture in Liverpool: Popular Music in the Making.* Clarendon Press / Oxford University Press, Oxford; New York, 1991.

[CS20] Samuel Cameron and Hendrik Sonnabend. Pricing the groove: Hedonic equation estimates for rare vinyl records. *Applied Economics*, 52(50):5516–5530, 2020.

[dis25] Discogs data dump, april 2025. https://discogs-data-dumps.s3.us-west-2.amazonaws.com/index.html?prefix=data/2025/, April 2025. Monthly release, Creative Commons CC0 license.

[DP01] Gilles Duranton and Diego Puga. Nursery cities: Urban diversity, process innovation, and the life cycle of products. *American Economic Review*, 91(5):1454–1477, December 2001.

[EP19] Jeff Ens and Philippe Pasquier. Quantifying musical style: Ranking symbolic music based on similarity to a style. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[Esp] Christopher R. Esposito. The geography of breakthrough invention in the united states over the 20th century. 52(7):104810.

[FAM17] Richard Florida, Patrick Adler, and Charlotta Mellander. The city as innovation machine. *Regional Studies*, 51(1):86–96, 2017.

[FJ10] Richard Florida and Scott Jackson. Sonic city: The evolving economic geography of the music industry. *Journal of Planning Education and Research*, 29(3):310–321, 2010. Available at SSRN: https://ssrn.com/abstract=2175348.

[FW] Michael Fritsch and Michael Wyrwich. Does successful innovation require large urban areas? germany as a counterexample. 97(3):284–308. Publisher: Routledge _eprint: https://doi.org/10.1080/00130095.2021.1920391.

[GKB21] Lukas Gienapp, Clara Kruckenberg, and Manuel Burghardt. Topological properties of music collaboration networks: The case of jazz and hip hop. *Digital Humanities Quarterly*, 15(1), 2021. Revision Note: A description of network density has been corrected. The previous version remains available.

[GN12] Isin Guler and Atul Nerkar. The impact of global and local cohesion on innovation in the pharmaceutical industry. *Strategic Management Journal*, 33(5):535–549, 2012.

[HAQ18] Desislava Hristova, Luca M. Aiello, and Daniele Quercia. The new urban success: How culture pays. *Frontiers in Physics*, 6, April 2018.

[HSS08] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[JTH93] Adam B Jaffe, Manuel Trajtenberg, and Rebecca Henderson. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics*, 108(3):577–598, 1993.

[KP12] Nathaniel Vaughn Kelso and Tom Patterson. World urban areas, landscan, 1:10 million (2012). http://www.naturalearthdata.com, 2012. Made with Natural Earth.

[KPSS12] Leonid Kogan, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman. Technological innovation, resource allocation, and growth. Technical Report No. 17769, National Bureau of Economic Research, January 2012. Revised May 2016.

[KPST21] Bryan Kelly, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3):303–320, 2021.

[Li] Tianyu Li. Modeling the local geography of country music concerts in u.s. urban areas: insights from big data analysis of live music events. 2(1):4.

[LMC13] Conrad Lee, Aaron F. McDaid, and Pádraig Cunningham. Measuring the significance of the geographic flow of music. *arXiv preprint arXiv:1301.5586*, 2013. Published January 2013.

[LRP14] N. Lee and A. Rodríguez-Pose. Creativity, cities, and innovation. *Environment and Planning A*, 46(5):1139–1159, 2014.

[Mak17] Teemu Makkonen. North from here: the collaboration networks of finnish metal music genre superstars. *Creative Industries Journal*, 10(2):104–118, 2017.

[MHS+23] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, Chris Cundy, Ziyuan Li, Rui Zhu, and Ni Lao. On the opportunities and challenges of foundation models for geospatial artificial intelligence, 2023.

[Mor] Stefano Moroni. Urban density after jane jacobs: the crucial role of diversity and emergence. 3(1):13.

[NFH14] Önder Nomaler, Koen Frenken, and Gaston Heimeriks. On scaling of scientific knowledge production in u.s. metropolitan areas. *PLoS ONE*, 9(10):e110805, 2014.

[Obend] J. Wolf Oberholtzer. discograph: Social graphing for the discogs database. GitHub repository, n.d. Retrieved August 9, 2025.

[SMT+19] Evan Sheehan, Chenlin Meng, Matthew Tan, Burak Uzkent, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2698–2706, New York, NY, USA, 2019. Association for Computing Machinery.

[SSC21] Marc Serra-Peralta, Joan Serrà, and Álvaro Corral. Heaps' law and vocabulary richness in the history of classical music harmony. *EPJ Data Science*, 2021.

[Swi] Marco Swiniartzki. Why florida?: Regional conditions and further development of the "florida death metal" scene and the local public response (1984–1994). 33(3):168–193.

[Tat17] Andy Tattersall. Disentangling the academic web: what might have been learnt from discogs and imdb, February 2017.

[TJ20] Timothy J. Tsai and Kevin Ji. Composer style classification of piano sheet music images using language model pretraining. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[Uzz97] Brian Uzzi. Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative Science Quarterly*, 42(1):35–67, 1997.

[VvN16] Marc Verboord and Sharon van Noord. The online place of popular music: Exploring the impact of geography and social media on pop artists' mainstream media attention. *Popular Communication*, 14(2):59–72, 2016.

[Wat12] Allan Watson. The world according to iTunes: mapping urban networks of music production. 1 2012.

[YDV24] David Yanagizawa-Drott and Hans-Joachim Voth. Image(s). Revise & Resubmit, *Quarterly Journal of Economics*, July 2024.

[ZL] Wenhao Zhou and Hailin Li. R&d team network configurations, knowledge diversity and breakthrough innovation: a combined effect framework. _eprint: https://www.emerald.com/ejim/article-pdf/doi/10.1108/EJIM-11-2023-1004/9759516/ejim-11-2023-1004.pdf.

A Appendix

A.1 Example of Discogs Data

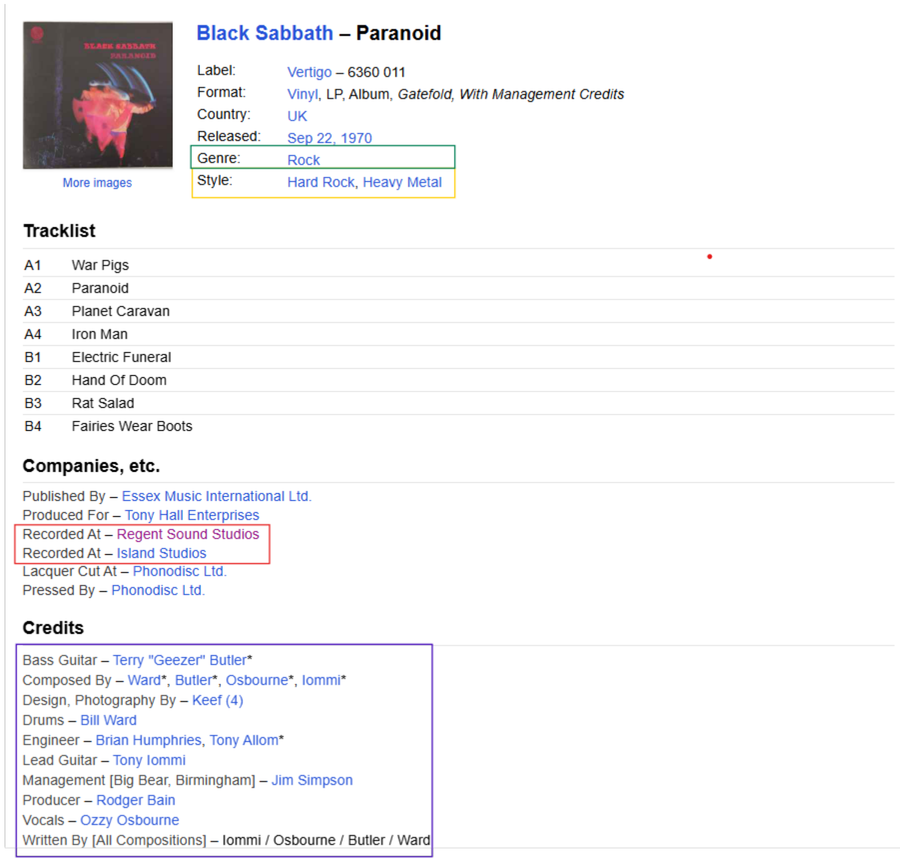

Figure 15: This figure shows the liner notes contained in the 1970 UK LP for Black Sabbath's Paranoid as it appears in the Discogs UI. The genre and styles are identified in green and yellow respectively. Studios, which are linked to the labels page, are identified in red. Lastly, the personnel information, linked to the artists database, is identified in blue. Note the blue text in UI - this indicates where the musician, tag, or recording studio links to a unique identifier with further information.

A.2 Genre Coverage in Sample

Genre	Count in Masters Database	Count in Sample	Percent Coverage
Classical	147,979	27,887	0.188
Jazz	192,362	34,934	0.182
Blues	51,132	8,611	0.168
Rock	743,936	124,109	0.167
Reggae	62,127	8,639	0.139
Funk / Soul	169,103	21,216	0.125
Stage & Screen	69,149	7,823	0.113
Folk, World, & Country	313,258	31,361	0.100
Pop	496,460	48,403	0.097
Hip Hop	115,121	10,645	0.092
Latin	102,532	7,308	0.071
Non-Music	41,359	2,919	0.071
Children's	24,052	1,638	0.068
Brass & Military	8,179	516	0.063
Electronic	667,341	33,576	0.050

Table 7: Genre representation in sample compared to Discogs Masters database.

A.3 Full Table of Average Innovation Scores (Across All Periods)

City	Score
Tampa	2.392
Newcastle	1.847
Leeds	1.842
Auckland	1.749
Manchester	1.741
Evanston	1.686
Liverpool	1.685
Düsseldorf	1.678
Alexandria	1.675
Chicago	1.669
Sheffield	1.581
Edinburgh	1.546
Long Beach	1.541
Dortmund	1.534
Kingston	1.504
Seattle	1.484
Philadelphia	1.472
Brussels	1.448
San Jose	1.440
Goteborg	1.415
San Mateo	1.400
Bristol	1.374
Irvine	1.373
Newark	1.367
Essen	1.363
Turin	1.355
Cleveland	1.350
Washington, D.C.	1.349
Milan	1.347
Melbourne	1.345
Dallas	1.343
Aarhus	1.342
Paterson	1.340
Malmo	1.335
Verona	1.318

City	Score
Glasgow	1.315
Houston	1.312
Sydney	1.303
San Francisco	1.295
Austin	1.280
Oakland	1.275
Detroit	1.255
New York	1.245
London	1.239
Berkeley	1.231
Vatican City	1.209
Antwerpen	1.202
Dublin	1.201
Vancouver	1.199
Portland	1.198
Hamburg	1.187
Birmingham	1.185
Toronto	1.179
Oslo	1.179
Memphis	1.163
Athens	1.159
Boston	1.147
Koln	1.147
Los Angeles	1.139
Bologna	1.138
Atlanta	1.131
Madrid	1.112
Zürich	1.109
New Orleans	1.106
Warsaw	1.100
San Diego	1.095
Tokyo	1.093
Helsinki	1.091
Copenhagen	1.086
Hannover	1.085
Pasadena	1.081

City	Score
Rome	1.081
Amsterdam	1.080
Minneapolis	1.068
Barcelona	1.059
Stockholm	1.051
Paris	1.048
Osaka	1.046
Rotterdam	1.046
Montreal	1.043
Berlin	1.034
Cincinnati	1.027
Kyoto	1.023
Miami	1.021
Nashville	1.021
Budapest	1.018
Frankfurt	1.015
Sao Paolo	1.012
Versailles	1.005
Stamford	0.996
Buenos Aires	0.992
Haarlem	0.986
The Hague	0.977
Ljubljana	0.972
Miami Beach	0.966
Stuttgart	0.966
Prague	0.960
Bucharest	0.934
Munich	0.924
Rio de Janeiro	0.877
Leipzig	0.874
Dresden	0.865
Geneva	0.806
Vienna	0.747

Table 8: Average Innovation Scores by City, Entire Dataset

A.4 Average Innovation Scores - First Releases Only

Rank	City	Avg. First Release Taddy Score
1	Seattle	4.9404
2	Detroit	3.1735
3	Miami	2.3344
4	Chicago	2.2766
5	Vancouver	2.2447
6	Manchester	2.1720
7	Dublin	2.1355
8	Verona	2.1111
9	Memphis	2.0819
10	Cleveland	2.0712

Table 9: Top 10 cities by Average Innovation of First Releases per Artist

A.5 List of Satellite Cities Excluded in Robustness Check

Satellite City	**Associated Metropolis**
Paterson, New Jersey	New York City
Newark, New Jersey	New York City
Evanston, Illinois	Chicago
Pasadena, California	Los Angeles
Irvine, California	Los Angeles
Alexandria, Virginia	Washington DC
Versailles, France	Paris
Oakland, California	San Francisco
San Jose, California	San Francisco
San Mateo, California	San Francisco
Berkeley, California	San Francisco
Oakland, California	San Francisco

Table 10: Satellite cities excluded from robustness checks.

A.6 Regressions with Satellite Cities Excluded

Table 11: Effect of Total Musical Output on Innovation

	Avg. Innovation Score of First Releases (log)	Avg. Innovation Score (log)	Innovation Score 75th percentile (log)	Innovation Score 95th percentile (log)
Lagged log(Albums)	-0.101***	-0.119***	-0.0746***	-0.190***
	(0.0246)	(0.0148)	(0.0149)	(0.0230)
Constant	5.310	-2.270	-0.203	3.925
	(4.871)	(1.760)	(1.769)	(2.734)
Genre Controls	Yes	Yes	Yes	Yes
Observations	330	353	353	353
R-squared	0.179	0.289	0.187	0.279

Standard errors in parentheses

Fixed effects included

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$