Mining and Forecasting Career Trajectories of Music Artists

Anonymous Author(s)

ABSTRACT

Many musicians, from up-and-comers to established artists, rely heavily on live performances to promote and disseminate their music. Furthermore, owing to the global slowdown of record sales over the past two decades, live performances have also become an important revenue source for musicians. To promote their concerts, artist often upload their tour dates to online platforms such as *Songkick* for others to see. In this article, we first present a new dataset we constructed by cross-referencing data from *Songkick* and *Discogs*, another web site containing highly granular information about music artists. We then demonstrate how this dataset can be used to mine and predict important career milestones for the musicians, such as signing by a major label, or playing at a high profile venue. We also perform a temporal analysis of the bipartite artist-venue graph, and demonstrate that high centrality on this graph is correlated with success. Our work contributes to the emerging field of *Science of Success* and shows how digital traces from online platforms can reveal and predict patterns of success in the offline world.

CCS CONCEPTS

•Information systems → Data mining; •Theory of computation → Social networks; •Computing methodologies → Machine learning approaches; Network science;

KEYWORDS

networks, art and music, multidisciplinary topics and applications

ACM Reference format:
Anonymous Author(s). 2018. Mining and Forecasting Career Trajectories of Music Artists. In *Proceedings of ACM Conference on Hypertext and Social Media, Baltimore, Maryland USA, 9-12 July 2018 (Hypertext'18),* 9 pages.
DOI: 10.475/123_4

1 INTRODUCTION

Live performances are a crucial part of the life of a music artist. According to a recent industry report [1], the revenues from live performances in the US have grown from $8.72B in 2012 to $9.94B in 2016, and are projected to almost reach $12B by 2022. A recent study discovered a connection between live events and increased digital listenership (which is the second highest source of income for a band after live performances). In light of this, it becomes increasingly more important for artists to be able to understand what milestones matter to accomplish the dream of a professional career: playing at top venues goes hand-in-hand with getting more digital listeners, which in turn may increase their likelihood of being signed with major music labels.

In this work, we aim to determine whether it is possible to model and predict these career trajectories under the emerging framework of *Science of Success* [8]: recent work studying how careers in different fields, as well as individual and team success, can be predicted early by leveraging records of performance from digital traces. This data-driven framework has been applied to domains as diverse as education and academia [16, 27], sports [6, 7, 33], social media [2, 11, 19, 28], and even the entertainment industry [23, 26].

In light of these promising results, we pose the question: is it possible to find open data to understand and forecast careers and success is the music industry? Fortunately, in recent years, bands have turned to online platforms like Facebook, Reddit, etc., as well as newborn websites, to advertise their concerts. To accommodate the increasing demand of music artists to get their message out to their fans, specialized sites like *Songkick* and *Discogs* have sprung up to create centralized repositories of music events and music artists. These sites contain rich metadata about the artists themselves as well as the concerts they perform. They allow the artists to attract interests in their concerts. Indirectly, this goldmine also allows researchers to model the music industry dynamics.

Research agenda

In this paper, we are interested in the problem of characterizing and understanding the career trajectories of the artists across different genres.

Toward this goal, we analyze a large-scale longitudinal data of musical events occurring at various venues worldwide.

Specifically, we address the following research questions:

(1) Is the choice of venues where an artist performs correlated with the eventual success of that artist (for a given definition of success)? And, if that is the case, can we leverage those correlations to forecast success?
(2) Can we predict which venues an artist/band will perform based on the history of his/her/their past performances?
(3) How do we measure the relative importance of performances in specific venues and their impact on career trajectories, and how do we jointly characterize *influential* artists and venues?

Contributions of this work

Our main contributions are summarized as follows:

- We construct and present a new dataset by collecting all of the artists and concerts from the *Songkick* platform, and supplement this dataset with information from *Discogs*, which contains more granular details about the artists—such as their discographies.
- We define a measure of success based on whether an artist has signed a contract with one of the major music record

[1] https://www.statista.com/statistics/491884/live-music-revenue-usa/

Hypertext'18, Baltimore, Maryland USA

DOI: 10.475/123_4

labels, and propose a forecasting task to differentiate between career trajectories of artist based on this measure of success.
- We demonstrate the viability of forecasting future performances of artists, and therefore their success, based on the history of past performances.
- We propose a centrality measure suited for the bipartite artist-venue network and demonstrate that it correlates strongly with the venue reputation.

The rest of the paper is organized as follows. After describing related work in Section 2, we describe the dataset in Section 3 and provide its basic statistics in Section 4. In Section 5 we define three related tasks - forecasting artists success, predicting future events by artist at specific venues, and identifying influential artists and venues - describe our approach for addressing those tasks, and present results. We conclude the paper by summarizing our main findings in Section 6.

2 RELATED WORK

We break the discussion of work related to this paper into two sections. We first introduce notable data-driven contributions to the emerging area of *Science of Success*, and contextualize our work within this framework. Then, we discuss related literature on forecasting methods, in particular at the intersection between network science and machine learning (i.e., network-based predictions), specifically discussing some recent prediction methods for bipartite networks.

Science of Success

Quantifying and forecasting success refers to the broader body of work that attempts to discover the patterns and performance trajectories that correlate with certain desirable outcomes: from forecasting highly-cited academic authors and papers [15, 31] to predicting future nobel prize winners [20], from uncovering successful fund-raising campaigns [21], to early identifying the next top model [23], or movie box office hit [10], the new field of *Science of Success* brings a strong data-driven perspective on applied forecasting problems set in the real world.

In this work, forecasting success is operationalized as predicting the artists or bands that are going to be signed by a major music recording label. To the best of our knowledge, this is a novel formulation that has not been presented in the literature before. However, in a recent paper, Rossetti and coauthors [26] looked at the popularity of music artists on digital delivery platforms like Last.fm, and formulated a forecasting problem to predict new song hits from the early adoption patterns of music listeners. Other work in the arts behavioral literature analyzed career trajectories of music artists [34], finding a correlation between the size of artists' social networks and their probability to succeed in the industry.

Network Forecasting Methods

From a methodological perspective, our work is rooted on a blend of machine learning and network science techniques. We focus in particular on a broad class of problems often referred to as *link mining* (a.k.a. *link prediction*). Link mining is the problem of discovering new (unforeseen) edges in a graph. Typical possible applications are either network reconstruction [9, 12], or modeling the evolution of a network [4, 14, 32]. One common operationalization of link prediction is finding pairs of nodes that have high probability of being connected. This often translates into measuring node similarities, as mentioned by Liben-Nowell and Kleinberg [18]. However, other authors [17] noted that using traditional link prediction on bipartite graphs is not straightforward and often produces counterintuitive results. In order to address this shortcoming, some authors proposed modified similarity metrics [17, 18], or used techniques from recommender systems, such as low-rank matrix factorization and collaborative filtering [1, 5], and supervised learning approaches [3, 24]. We follow the example of those authors and use collaborative filtering and recommender systems inspired methods to perform link prediction for our task. In the results section, we will show how to leverage *BiRank* [13]—a modification to the *PageRank* [22] algorithm that tunes it towards bipartite graphs—to measure and predict the popularity of the artists and venues.

3 DATASET

Songkick[2] is a concert-discovery platform that aims to link fans to artists' events. It contains information about over 6 million concerts (and other music events like festivals), the artist(s) that perform at each event, and the venue where each event takes place. The "gigography" of an artist is the term that Songkick uses to refer to all of that artist's events.

Songkick data can be accessed through their website or via their API, which allows querying any artist's gigography. Songkick is our main repository of information for musical events.

Discogs[3] is a music database that contains cross-referenced discographies of artists and labels. Each recording, artist, or label in Discogs can be uniquely identified by their IDs. Discogs provides separate data dumps[4] for artists, labels, and recordings. We used recordings data dump from May 1, 2017 to obtain artist and label IDs associated with each release. This data dump contains more than 8 million recordings. Most of the recordings have information about their release dates, and thus allow tracking the history of releases with different labels for each artist.

3.1 Data Collection

Songkick does not provide a lookup directory of artists, nor there is a direct mechanism to get all the gigographies. For getting Songkick artist IDs we queried artist names present in Discogs' recordings data dump. As a result, all of the artists in our dataset have at least one recording on Discogs, which can be either self-recorded or recorded under a contract with a music label. This strategy avoids the introduction of a bias towards artists that did not publish any recordings, which are therefore excluded from our analysis.

The Songkick API call returns a list of possibly relevant artists, allowing for some inexact name matching. We processed the API output to retain data on artists that exactly matched the Discogs artist name.

From this name match we obtained artist IDs, and used them for another round of API calls, to get the gigographies of each artist.

[2] https://www.songkick.com/
[3] https://www.discogs.com/
[4] https://data.discogs.com/

For each concert in the gigography, we extracted the following information: ID, date, city, country, state (if applicable), latitude and longitude of the venue, venue ID and venue name, name of the event and its popularity score as calculated by Songkick.

For every event there is information about billing for each artist, i.e., whether that artist was a headliner or a support artist at the concert. However, we did not consider headliners and support artists separately in the analysis presented further.

The collected data was organized into separate artist, event, and venue data frames. Each artist is indexed by its Songkick and Discogs IDs. Venues and events are indexed by their Songkick IDs. There are also several lists of cross-references: mapping venues to the events that happened there, and events to the venues where they took place. A similar mapping is available for events and artists, and releases and artists.

3.2 Data Preprocessing

Due to the fact that the goal of Songkick is connecting fans to their favorite artists through concerts, the platform puts less relevance on events that occurred prior to their inception. Songkick was founded in 2007 and there is a noticeable increase in the number of artists that have their earliest concerts recorded on Songkick in 2007 or later (see Figure 2 in the next section). For the sake of data completeness, we focus only on artists that have their first record of performance in 2007 or later. By doing so, we aim to retain only the artists who used Songkick to inform their fans about upcoming events, thus avoiding the use of possibly incorrect backdated data.

In this paper we consider an artist that has one or more recordings with one of the major labels(a.k.a., "Big Three"/Four/Five/Six[5]), or their subsidiaries, "successful". Conveniently, each music record label in Discogs has information about its sub-labels and its parent label, if such exist. This allowed tracking all subsidiaries of the major labels. We assume the first time an artist releases a recording with such a label to be the change point in their career. We are interested in researching the trajectory of artists before the change point and ideally being able to forecast the change point.

As one of the main contributions in our paper is to the Science of Success, we wanted to make sure that we have enough data about successful artists in the early stage of their career. Thus, in a final preprocessing step, we removed every artist and venue that has less than 10 concerts associated with them before the change point. This also takes care of venues that may have been used for occasional events, or artists whose career terminated immediately after its start.

4 STATISTICS

In the following we provide some statistical analyses of our dataset. The dataset contains 645,507 concerts, 13,912 artists, and 11,428 venues, collected for the time frame between 2007 and 2017. Artists in the dataset are associated with 39,641 distinct record labels, 286 of which are major labels, or their direct subsidiaries. One condition to be labeled as a "successful' artist in our study is to have recorded at least one album under any of these 286 recording labels.

Figure 1 depicts distributions of the number of concerts per artist and number of concerts per venue. Both distributions are

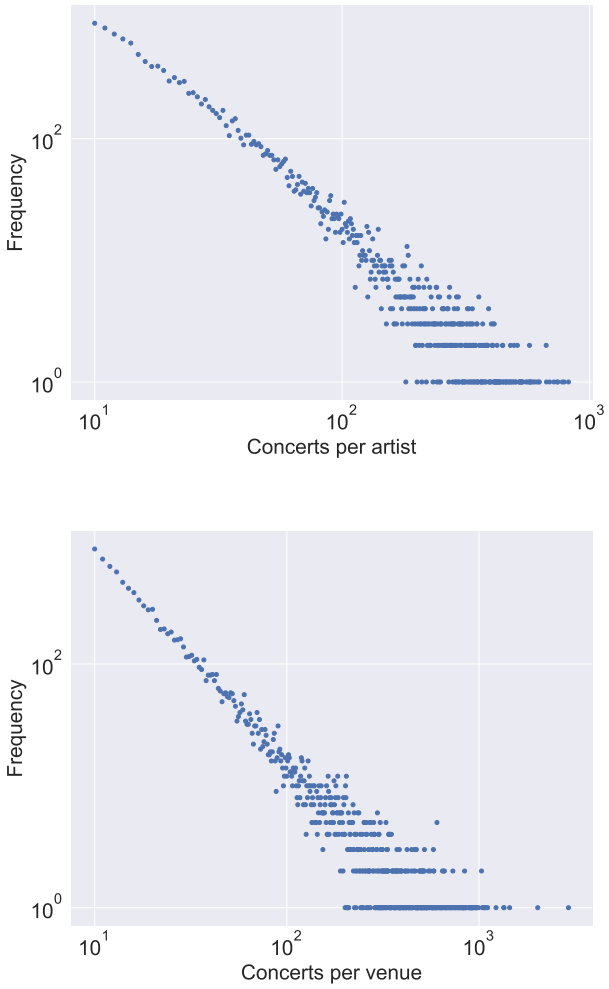

Figure 1: Heavy-tailed distributions of the number of concerts per artist (upper panel) and per venue (lower panel), respectively.

very broad and heavy tailed, with few active artists and venues hosting many events, and a very large set of artists and venues associated with very few events.

In Figure 2 we show the dynamics of the number of events and number of artists from 1987 to 2017. As already mentioned, there is significant increase in the number of artists that have their earliest concerts recorded on Songkick in 2007 or later. From the Figure 2 it can be seen that the total number of concerts per year peaked in 2010.

Next, we look at the geographic distribution of venues in the dataset. There are 63 different countries with at least on event, which for the mostpart are in North America and Europe. Almost half of all venues are located in the United States, where also more than half of all concerts happened. The second highest numbers for both number of concerts and venues are in the UK. Figure 3

[5] https://en.wikipedia.org/wiki/Record_label#Major_labels

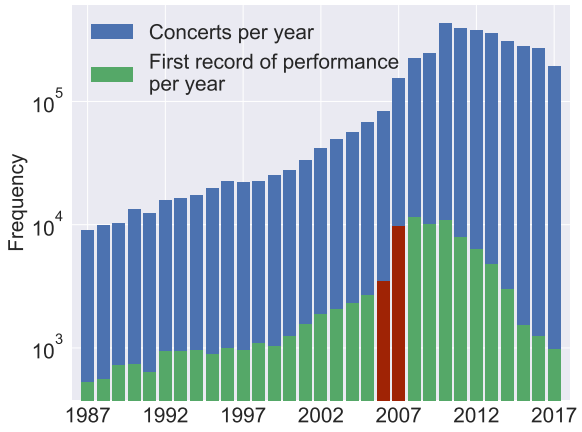

Figure 2: Number of concerts that are present in the dataset and the number of artists that first appear on Songkick in a given year. The red bars illustrate the sudden big change from 2006 to 2007 in the number of artists that first appeared on Songkick in those years. We believe this can be explained by the fact that Songkick was founded in 2007. Data before 1987 are very limited thus not included in this illustration.

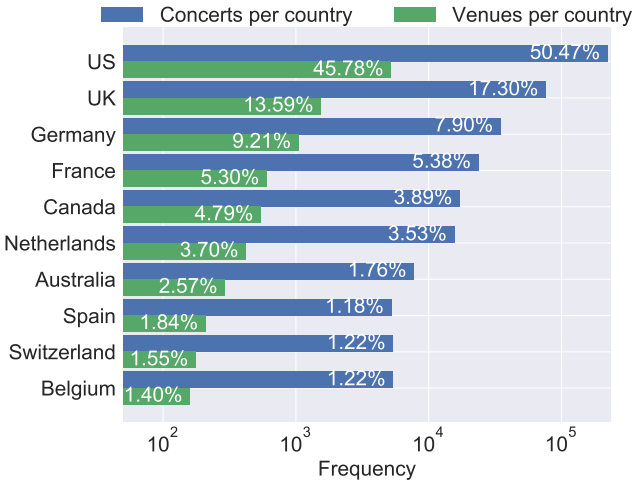

Figure 3: Log-scale distribution of concert frequencies in (i) the top 10 most active countries, and (ii) the number of distinct venues in those countries. A disproportionate preference toward English-speaking countries can be observed in the Songkick data, with United States, United Kingdom, and Australia cumulatively accounting for nearly 70% of the total events, and over 60% of the total venues.

Table 1: Some of the most frequent n-grams extracted from sequences of artists' performances. Double-sided arrows indicate that these routes are frequently found in the data in both directions.

Frequent routes that artists follow
San Diego ↔ Los Angeles ↔ SF Bay Area ↔ Portland ↔ Seattle
Portland ↔ Seattle ↔ Boise ↔ Salt Lake City ↔ Denver
Chicago ↔ Toronto ↔ Montreal ↔ Boston/Cambridge ↔ New York
Washington ↔ Philadelphia ↔ New York ↔ Boston/Cambridge
London ↔ Birmingham ↔ Manchester ↔ Glasgow
Brisbane ↔ Sydney ↔ Melbourne ↔ Adelaide
Austin ↔ Houston ↔ New Orleans ↔ Atlanta

demonstrates distribution of the venues and concerts at the top 10 most frequently occurring countries.

If we look at more granular information about geolocation of artists' performances we can get an insight on actual spatial trajectories of artists. Particularly, we can look for frequent subsequences among the sequences of performances that artists had. As displayed in Table 1, n-grams of length 4 and 5 show some frequent routes of cities, that artists take while touring. Following the distribution of the venues and concerts in the dataset, most frequent routes mostly include US cities. The frequent routes as demonstrated in the Table 1 contain clear patterns of artists performing in big cities on their way while travelling from North to South or from East to West, etc.

5 ANALYSIS AND RESULTS

To better illustrate the idea that the music artist career trajectory can be predicted from artist-venue interactions we formulated the following 3 tasks, discussed next:

- Task 1: Forecasting artist success;
- Task 2: Event prediction;
- Task 3: Joint discovery of influential artists and venues.

In the next subsections, we describe each of those tasks in more details, elaborate on our approach for addressing them, and present our results.

5.1 Task 1: Forecasting Artist Success

Due to the nature of the partnership between artists and record companies, the bigger the recording label the more resources and opportunities it has to offer for its artists. Artists, nurtured by labels, have the chance to develop their sound, their craft, and their careers. Besides, record companies facilitate introductions to world-class producers, writers, and other performers, which can determine careers and bring huge rewards.

The recording industry has been marked by concentration and centralization for a while now. During the phase of consolidation in 1970s, most of the major labels were acquired by very few umbrella

corporations or music groups. The Beatles, Frank Sinatra, Pink Floyd and even Maria Callas found prominence through those major record labels.

From 1988 till 2012 the number of major record companies has decreased from six to three, as some of them got absorbed by the others. The remaining three major music groups, or the *Big Three* (Sony BMG, Universal Music Group, and Warner Music Group), have held a large share of the world music production since 2012.

Because of the influence and patronizing that the major labels provide, we consider artists that have a recording with either the parent major label, or one of its direct subsidiaries, as "successful". We set to see if the rise to success can be predicted from a sequence of performances. Our goal in this task is therefore to identify successful artists from their career trajectories.

Ideally, we want to be able to identify such artists in a post-hoc manner. In other words, we want to detect the change that will lead to a release with a major label before the release itself happens. In the following discussion we refer to this task as *forecasting*.

However, we also consider the simpler task of discriminating artists that are already successful in our setup from the ones that are not. We refer to this task as *prediction*.

5.1.1 Experimental Setting. For both forecasting and prediction tasks we used the *affiliation matrix* of artists and venues. In such an affiliation matrix, an artist is represented as a bag-of-words vector over the venues where the artist has performed. We used those vectors as features for the prediction and forecasting tasks.

In the forecasting task, we did not include any concert that happened after an artist released their first recording with a major music label. However, for the prediction task we included those performances too.

The classification labels (successful or not) were obtained by iterating over all the music labels that each artist has ever recorded with (this information was obtained from Discogs). If among these music labels there are either major ones or one of their subsidiaries, we assume that the artist was successful and label it as a positive instance—negative otherwise.

As a result of the procedure above, we labeled about 500 artists as successful, which is 3.6% of the total. It is worth noting that our labeling procedure yields an highly-unbalanced dataset, where the positive instances (successful artists) are very infrequent: this is in line with the commonsense notion of popularity in the music industry, where musicians and bands that thrive with a professional career are exceptionally rare.

5.1.2 Metrics. A natural choice for evaluating a success forecasting or prediction task is classification accuracy. However, due to high imbalance in the data, we need metrics that are more sensitive and account for under-represented classes. Such metrics are Precision, Recall and F1 score, as well as ROC AUC score, which we used for evaluation.

5.1.3 Learning Models and Configuration. For Task 1, we defined three simple models described next, and used them to carry out the forecasting and predictions exercises.

Baseline: We used the "True" labels for every artist as a baseline. Given the unbalance issue discussed above, this baseline will have very low Precision (equal to the frequency of True labels, i.e., 3.6%), perfect Recall (all successful artists will be label as such), and a ROC AUC score of 50%, equivalent to random guessing.

Logistic Regression: As a base classifier in both prediction and forecasting experiments we used Logistic Regression from the scikit-learn library [25]. We used L_2 norm for regularization, and tuned one parameter, i.e., the inverse of regularization strength C.

SVD: Since the affiliation matrix we use has over 99% sparsity (percentage of zero entries), dimensionality reduction techniques could yield prediction performance improvements by transforming sparse data into dense. We performed dimensionality reduction using Singular Value Decomposition (SVD). We report the results for reducing the space of venues to 500 components.

CorEx: Covariance matrices and hence eigenvalues are massively amplified in high-dimensional data, where the number of samples is comparable to the number of features. Thus, besides using a simple SVD, we also wanted to employ a technique that is designed for dealing with high-dimensional data. We thus also used Total Correlation Explanation [29, 30] (CorEx) for dimensionality reduction.[6] We report results for 350 latent components.

For each model, we performed hyper-parameter tuning via grid search with 3-fold cross validation on the training set. The results reported are obtained by using cross-validated average over 3 different train-test splits in 80-20 ratio.

5.1.4 Task Summary. The results for this task are presented in Table 2. Even a simple logistic regression achieves 13% precision on the forecasting and prediction tasks. For logistic regression on full data the results of prediction task are marginally better than those of forecasting task, but on reduced dimensions we were able to achieve 10% higher precision and 0.1 higher F1-score for prediction than for forecasting. Both dimensionality reduction techniques give better precision and F1-score on both tasks, while having comparable AUC score. The results for CorEx are slightly better than those of SVD, which proves our point that this task is better tackled with specialized methods for high-dimensional sparse data.

The better performance of our methods on the prediction task shows that there is a difference in distributions of artist performances before and after they record their first album with a major recording label. This suggests the existence of a change point in the career of those artists that is caused by recording with a major label, which corroborates our notion of artist's success. We expect that employing more sophisticated models for discovering that change point would give better forecasting results.

5.2 Task 2: Event Prediction

Besides artist career trajectories, we are also interested in the overall dynamics of the network, where both venues and artists evolve and their influence changes as a result of constant interactions between venues and artists.

To see if we can explain part of those interactions, we formulate the artist-venue link prediction task. As in the previous artist success task, we consider here the same two configurations—*forecasting* and *prediction*. For this task we used the same affiliation network as in the previous task.

In the previous task prediction experiments were performed to test whether or not our suggested definition of success is viable. For

[6] Implementation available here: https://github.com/gregversteeg/corex_topic

Table 2: Precision (P), Recall (R), F1-score and AUC for artist success forecasting (FCST) and prediction (PRED) tasks. We show results of logistic regression on full data (FCST/PRED LR), and with reduced dimensions, using SVD (FCST/PRED LR+SVD), and CorEx (FCST/PRED LR+CorEx), respectively.

Task	Model	P	R	F1	AUC
	Baseline	0.03	1.00	0.07	0.50
FCST	LR	0.13	0.62	0.21	0.72
FCST	LR+SVD	0.14	0.51	0.22	0.69
FCST	LR+CorEx	0.18	0.49	0.26	0.70
PRED	LR	0.13	0.67	0.22	0.74
PRED	LR+SVD	0.23	0.58	0.33	0.76
PRED	LR+CorEx	0.27	0.57	0.36	0.75

($artist - venue$) link mining task, however, we exercise prediction alongside to the forecasting to test for possible major temporal shifts in artists' behavior.

5.2.1 Experimental Setting. In the forecasting task, we looked for new ($artist, venue$) links based on the history of old ones. In particular, we used all performances from 2007 to 2015 as history (i.e., training data), and the performances in 2016 and 2017 as "future" (i.e., test set). We then went on and removed all artists that have less than 5 concerts either in train or test set. As a result we had 2,169 artists, 5,182 venues, 109,275 events in the training set and 42,035 events in the test set.

In the prediction task, instead, we randomly picked 20% of all links and hid them in the test data, using the remaining 80% for training purposes, similarly to a link prediction problem. We treated the links as multilinks based on the frequency of performances. Similarly to the forecasting task, we removed all artists that have less than 5 concerts either in train or test set. In this setup we ended up having 3,411 artists, 7,881 venues, 222,616 events in the training set and 50,440 events in the test set.

5.2.2 Metrics. We measured the performance on this task using Area Under the Receiver Operating Characteristic curve (ROC AUC). One of the main advantages of this metric is the fact that it operates on rankings and is calculated for a range of thresholds, rather than prediction classes. This allows us to interchangeably use simple recommender system objectives for venue prediction.

5.2.3 Learning Models and Configuration. For Task 2, we decided to adopt two baseline models and a simple matrix factorization technique, all described in the following.

Baselines: The baseline models that we employed for this task were designed to follow the intuition that venues convey some temporal patterns, which would allow to predict artists' career trajectories. One of the baseline is shared among the two settings, another is specific of the temporal prediction setup.

For the shared baseline, we used all the venues from the training set as prediction targets for possible venues in the test set. Besides it, for the temporal setup, we also tried using only the venues from the last two years worth of concerts—2014 and 2015—to predict the venues in 2016 and 2017. The difference in performances of these two baselines for temporal setup can indicate interesting trends in artist and venue changes over time. At the same time, using these baselines allow us to investigate whether or not there is a tendency to come back to play at venues where artists have played at previously.

Matrix factorization: Link mining in a bipartite graph can be naturally presented as a recommendation task. For each artist we have a list of "relevant" venues—the ones where the artist performed. Using methods for collaborative filtering we can find latent features or representations of venues that make them relevant for certain artists. Based on these hidden representations, we can then predict what are the venues that are most relevant for the artist. In contrast to traditional recommender system where the system only recommends items that user have not interacted with yet, in our case the artist may perform at both new venues and the old ones.

In this task, we used a simple yet popular collaborative filtering method based on matrix-factorization—Singular Value Decomposition (SVD). For SVD, we used cross-validated grid search on the number of hidden features to use—from 10 to 5000—and reported the result for 500. The results are very robust to variations of this parameter, and the observed dynamics are qualitatively unchanged.

5.2.4 Task Summary. The results for the venue prediction task are presented in Table 3. SVD outperforms our simple baselines in both prediction and forecasting tasks. The baseline including full training set performs better on the forecasting data. Our intuition was that the 2-year baseline could cover patterns that are temporally closer and hence more relevant to the testing period. However, experiments show that even if this is true, that does not lead to superior performance in terms of prediction quality. The fact that the baselines perform better than random, indicates that the habit of performing in the same set of venues is typical of music artists. Moreover, this tendency holds in both prediction and forecasting configurations, i.e. this habit describes both short and long-term preferences.

In this task we see SVD simplify the model and remove just enough noise to perform better under prediction configuration compared to forecasting configuration. Hence, similar to the artist success task, based on the results of these experiments we can come to conclusion that the difference in the distributions of the data in the prediction and forecasting setups can be captured by a model such an SVD. However, it can not be described or explained by simple assumptions that we made while constructing our baselines.

5.3 Task 3: Joint Discovery of Influential Artists and Venues

In the previous tasks, we have attempted to classify an artist as about to be signed or not about to be signed. In this task we will investigate whether we can identify top artists and venues automatically by mining their performances.

To measure the popularity of the artists and venues, we leverage BiRank [13]. This algorithm is a modification to the PageRank [22] algorithm that tunes it towards bipartite graphs. The algorithm iteratively identifies influential venues by observing which influential artists play at them. Simultaneously, it measures influential artists by measuring their frequency of playing at influential venues.

Table 3: Results for $(artist, venue)$ link prediction task, measured in Area Under Receiver Operating Characteristics curve (AUC). 2007-2015, 2014-2015 and 2007-2017 are the baselines using the training set, composed from events of the corresponding years, as a target prediction.

Task	Model	AUC
FCST	2007-2015	0.65
FCST	2014-2015	0.60
FCST	SVD	0.81
PRED	2007-2017	0.59
PRED	SVD	0.90

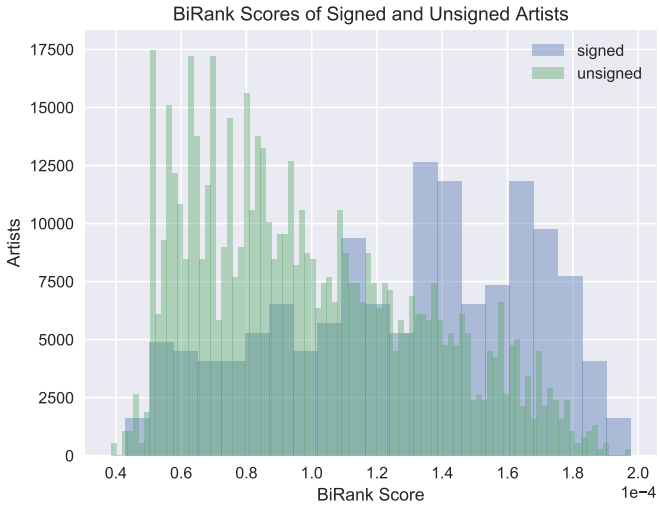

Figure 4: Histogram of signed and unsigned artists. Normalized to show relative frequency of BiRank scores.

Before running this algorithm, we set the initial ranking based upon the following measure:

$$g_i = \frac{\log(N_i + 1)}{\sum_{a \in \mathcal{A}} \log(N_a + 1)}, \tag{1}$$

where N_i measures the number of links to the node i, $\mathcal{A}$ is the set of artists in the dataset, and $i \in \mathcal{A}$. This constitutes the artist's initial score. Similarly, we compute:

$$g_j = \frac{\log(N_j + 1)}{\sum_{v \in \mathcal{V}} \log(N_v + 1)}, \tag{2}$$

where $\mathcal{V}$ is the set of venues and $j \in \mathcal{V}$. With this initial seed score, we proceed to run the BiRank algorithm to identify the most influential nodes in each set. Finally, it is important to note that there is a temporal weighting in the links. Each link in the adjacency matrix has a weight of δ^{2017-y_0}, where delta is the decay parameter (set to 0.85 in the experiments), and y_0 is the year of the first link. We subtract this number from 2017 as this is the most recent year in the dataset. This experimental setup closely resembles that of [13].

The results of this experiment can be seen in 4. These results seem to indicate promise for this method on our dataset. In the case of the venues, they correspond to some of the most popular venues in the world. As for the artists, the story is different. While they do not correspond to the most popular in terms of followers, these are the bands that have more performances in the dataset. However, a natural question regarding the dynamics of BiRank is how indicative it is of artist success. To measure this phenomenon, we plot the histogram of BiRank scores for both signed and unsigned artists. This can be seen in Figure 4, where we see that the signed artists tend to have a higher BiRank score than unsigned artists.

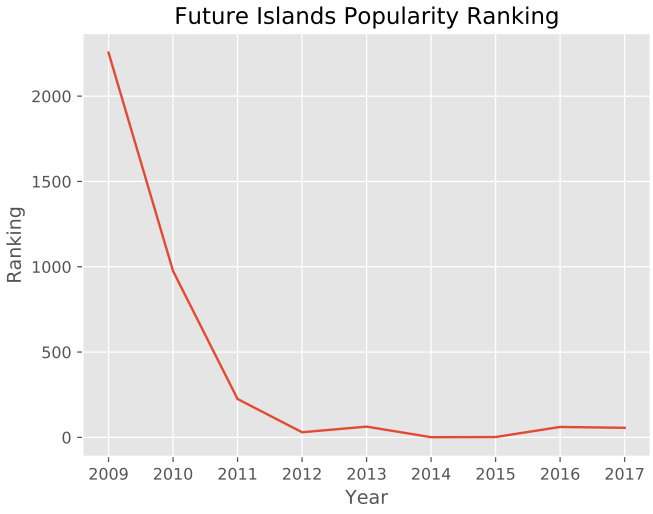

Figure 5: Trajectory of the group "Future Islands" through the lens of the BiRank score. The y-axis is the rank: lower is better. The BiRank score tracks the band's rise to popularity, culminating in the 2014 nomination of "breakthrough band of the year" by The Telegraph, suggesting that our framework can capture, and may predict, outstanding trajectories.

The BiRank scores can also be useful for measuring the trajectory of an artist. By calculating the BiRank scores as previously indicated every year, with a three year moving window, we can observe the ranking of artists at different points in time. An example of this phenomenon can be seen in Figure 5. This figure shows the BiRank ranking of the artist "Future Island" over time. We can see that their ranking begins around the 2,300 mark. Over the course of the next years, their ranking dramatically improves, peaking with them being the top artist according to BiRank in 2014. This is corroborated by The Telegraph naming them the "breakthrough band of the year."[7]

6 CONCLUSION

In this paper we presented a novel dataset of artists and their live performances from Songkick. We complemented that data by information collected from Discogs, which contains full history of their recordings and releases. The dataset can be used for a variety of tasks which we exemplified by performing success forecasting and event prediction.

[7] www.telegraph.co.uk/culture/music/music-festivals/10975049/Latitude-Festival-2014-Future-Islands-the-breakthrough-band-of-the-year.html

Table 4: The most influential nodes of each class identified by BiRank.

Rank	Artists	Venues
1	Frank Turner	The Observatory, Los Angeles, CA
2	Every Time I Die	The Masquerade, Atlanta, GA
3	Against Me!	The Bowery Ballroom, New York, NY
4	Reel Big Fish	Webster Hall, New York, NY
5	All Time Low	9:30 Club, Washington, DC
6	The Black Dahlia Murder	House of Blues, Boston / Cambridge, MA
7	Hatebreed	Theater of the Living Arts, Philadelphia, PA
8	Future Islands	The Middle East Downstairs, Boston / Cambridge, MA
9	Halestorm	Vienna Arena (Arena Wien), Vienna
10	Hawthorne Heights	Brudenell Social Club, Leeds

We proposed an operational definition of *success* - signing with a major label and/or their subsidiaries - and demonstrated that the event data contains useful information that can be leveraged to forecast artists' success with better than baseline accuracy. Similarly, we observed that by utilizing the underlying structure of this data, one can also predict whether an artist will have a concert in a particular venue. The performance of simple baseline models that we carried out in all three tasks indicates that much better results can be achieved with more carefully designed methods.

Finally, we illustrated how artist or venue influence can be measured based on analyzing a time-varying bipartite artist-venue graph. Specifically, we analyzed the evolution of the bipartite generalization of the Pagerank score, and demonstrated both qualitatively and quantitatively that its dynamics can be used to identify successful artists.

As future work, it will be interesting to perform more fine-grained analysis of all three tasks examined here. For instance, the results presented here were averaged across different genres. It is plausible, however, that analysis will yield (subtle) differences when conditioned on the genre. Similarly, our preliminary analysis of event sequence (as opposed to bag of word representation of events) yielded some interesting geographic patterns, which warrant further and more detailed studies. Finally, we note that despite its demonstrated usefulness, the dataset presented here is by no means perfect and might have some intrinsic biases, e.g., musicians might have varying incentives for joining platforms such as *Songkick* depending on the stage of their career. Identifying and potentially correcting for such biases is an important future task.

ACKNOWLEDGEMENTS

EF is partly supported by DARPA (grant no. D16AP00115). This project does not necessarily reflect the position/policy of the Government; no official endorsement should be inferred. Approved for public release; unlimited distribution.

REFERENCES

[1] Evrim Acar, Daniel M Dunlavy, and Tamara G Kolda. 2009. Link prediction on evolving data using matrix and tensor factorizations. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*. IEEE, 262–269.

[2] Roja Bandari, Sitaram Asur, and Bernardo A Huberman. 2012. The pulse of news in social media: Forecasting popularity. *ICWSM* 12 (2012), 26–33.

[3] Nesserine Benchettara, Rushed Kanawati, and Celine Rouveirol. 2010. Supervised machine learning applied to link prediction in bipartite social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*. IEEE, 326–330.

[4] Catherine A Bliss, Morgan R Frank, Christopher M Danforth, and Peter Sheridan Dodds. 2014. An evolutionary algorithm approach to link prediction in dynamic social networks. *Journal of Computational Science* 5, 5 (2014), 750–764.

[5] Krisztian Buza and Ilona Galambos. 2013. An application of link prediction in bipartite graphs: Personalized blog feedback prediction. In *8th Japanese-Hungarian Symposium on Discrete Mathematics and Its Applications June*. 4–7.

[6] Paolo Cintia, Luca Pappalardo, and Dino Pedreschi. 2013. " Engine Matters": A First Large Scale Data Driven Study on Cyclists' Performance. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*. IEEE, 147–153.

[7] Paolo Cintia, Salvatore Rinzivillo, and Luca Pappalardo. 2015. A network-based approach to evaluate the performance of football teams. In *Machine learning and data mining for sports analytics workshop, Porto, Portugal*.

[8] Aaron Clauset, Daniel B Larremore, and Roberta Sinatra. 2017. Data-driven predictions in the science of science. *Science* 355, 6324 (2017), 477–480.

[9] Aaron Clauset, Cristopher Moore, and Mark EJ Newman. 2008. Hierarchical structure and the prediction of missing links in networks. *Nature* 453, 7191 (2008), 98–101.

[10] Chrysanthos Dellarocas, Xiaoquan Michael Zhang, and Neveen F Awad. 2007. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive marketing* 21, 4 (2007), 23–45.

[11] Emilio Ferrara, Roberto Interdonato, and Andrea Tagarelli. 2014. Online popularity and topical interests through the lens of instagram. In *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, 24–34.

[12] Roger Guimerà and Marta Sales-Pardo. 2009. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences* 106, 52 (2009), 22073–22078.

[13] Xiangnan He, Ming Gao, Min-Yen Kan, and Dingxian Wang. 2017. Birank: Towards ranking on bipartite graphs. *IEEE Transactions on Knowledge and Data Engineering* 29, 1 (2017), 57–71.

[14] Hisashi Kashima and Naoki Abe. 2006. A parameterized probabilistic model of network evolution for supervised link prediction. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE, 340–349.

[15] Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. 2015. Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences* 112, 24 (2015), 7426–7431.

[16] Gregor Kennedy, Carleton Coffrin, Paula De Barba, and Linda Corrin. 2015. Predicting success: how learners' prior knowledge, skills and activities predict MOOC performance. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. ACM, 136–140.

[17] Jérôme Kunegis, Ernesto W De Luca, and Sahin Albayrak. 2010. The link prediction problem in bipartite networks. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems*. Springer, 380–389.

[18] David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology* 58, 7 (2007), 1019–1031.

[19] Zongyang Ma, Aixin Sun, and Gao Cong. 2013. On predicting the popularity of newly emerging hashtags in twitter. *Journal of the Association for Information Science and Technology* 64, 7 (2013), 1399–1410.

[20] Amin Mazloumian, Young-Ho Eom, Dirk Helbing, Sergi Lozano, and Santo Fortunato. 2011. How citation boosts promote scientific paradigm shifts and nobel prizes. *PloS one* 6, 5 (2011), e18975.

[21] Tanushree Mitra and Eric Gilbert. 2014. The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM,

49–61.

[22] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.

[23] Jaehyuk Park, Giovanni Luca Ciampaglia, and Emilio Ferrara. 2016. Style in the age of instagram: Predicting success within the fashion industry using social media. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 64–73.

[24] Milen Pavlov and Ryutaro Ichise. 2007. Finding experts by link prediction in co-authorship networks. In *Proceedings of the 2nd International Conference on Finding Experts on the Web with Semantics-Volume 290*. CEUR-WS. org, 42–55.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[26] Giulio Rossetti, Letizia Milli, Fosca Giannotti, and Dino Pedreschi. 2017. Forecasting success via early adoptions analysis: A data-driven study. *PloS one* 12, 12 (2017), e0189096.

[27] Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. 2016. Quantifying the evolution of individual scientific impact. *Science* 354, 6312 (2016), aaf5239.

[28] Gabor Szabo and Bernardo A Huberman. 2010. Predicting the popularity of online content. *Commun. ACM* 53, 8 (2010), 80–88.

[29] Greg Ver Steeg and Aram Galstyan. 2014. Discovering structure in high-dimensional data through correlation explanation. In *Advances in Neural Information Processing Systems*. 577–585.

[30] Greg Ver Steeg and Aram Galstyan. 2015. Maximally informative hierarchical representations of high-dimensional data. In *Artificial Intelligence and Statistics*. 1004–1012.

[31] Dashun Wang, Chaoming Song, and Albert-László Barabási. 2013. Quantifying long-term scientific impact. *Science* 342, 6154 (2013), 127–132.

[32] Linhong Zhu, Dong Guo, Junming Yin, Greg Ver Steeg, and Aram Galstyan. 2016. Scalable temporal latent space inference for link prediction in dynamic social networks. *IEEE Transactions on Knowledge and Data Engineering* 28, 10 (2016), 2765–2777.

[33] Claudia Zuber, Marc Zibung, and Achim Conzelmann. 2015. Motivational patterns as an instrument for predicting success in promising young football players. *Journal of sports sciences* 33, 2 (2015), 160–168.

[34] Koos Zwaan, Tom FM ter Bogt, and Quinten Raaijmakers. 2010. Career trajectories of Dutch pop musicians: A longitudinal study. *Journal of Vocational Behavior* 77, 1 (2010), 10–20.