Mining and Forecasting Career Trajectories of Music Artists

Shushan Arakelyan
USC Information Sciences Institute
shushan@isi.edu

Fred Morstatter
USC Information Sciences Institute
fredmors@isi.edu

Margaret Martin
USC Information Sciences Institute
mart586@usc.edu

Emilio Ferrara
USC Information Sciences Institute
ferrarae@isi.edu

Aram Galstyan
USC Information Sciences Institute
galstyan@isi.edu

ABSTRACT

Many musicians, from up-and-comers to established artists, rely heavily on performing live to promote and disseminate their music. To advertise live shows, artists often use concert discovery platforms that make it easier for their fans to track tour dates. In this paper, we ask whether digital traces of live performances generated on those platforms can be used to understand career trajectories of artists. First, we present a new dataset we constructed by cross-referencing data from such platforms. 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 music label, or performing at a certain venue. Finally, we perform a temporal analysis of the bipartite artist-venue graph, and demonstrate that high centrality on this graph is correlated with success.

CCS CONCEPTS

• **Information systems** → **Data mining**; *Web mining*; • **Networks** → **Online social networks**; • **Human-centered computing** → **Collaborative and social computing**; • **Computing methodologies** → *Machine learning approaches*; *Network science*;

KEYWORDS

networks, art and music, multidisciplinary topics and applications

ACM Reference Format:
Shushan Arakelyan, Fred Morstatter, Margaret Martin, Emilio Ferrara, and Aram Galstyan. 2018. Mining and Forecasting Career Trajectories of Music Artists. In *HT '18: 29th ACM Conference on Hypertext and Social Media, July 9–12, 2018, Baltimore, MD, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3209542.3209554

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 reach almost $12B by 2022. A recent study discovered a connection between live events and increased digital listenership [37] (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* [9, 14]: 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 [17, 21, 35], (e)sports [7, 8, 34, 39, 42], social media [3, 13, 24, 36], culture [2, 40], and even the entertainment industry [29, 33].

In light of these promising results, we pose the question: is it possible to find open data to understand and forecast careers and success in the music industry? 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 Problem

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)? If so, 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?

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

HT '18, July 9–12, 2018, Baltimore, MD, USA

ACM ISBN 978-1-4503-5427-1/18/07…$15.00
https://doi.org/10.1145/3209542.3209554

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.[2]
- We define a measure of success based on whether an artist has signed a contract with one of the major music record 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 artist 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

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 [20, 38] to predicting future Nobel Prize winners [25], from uncovering successful fund-raising campaigns [27], to early identifying the next top model [29], or movie box office hit [11]. The new field of *Science of Success* brings a strong data-driven perspective on applied forecasting problems set in the real world.

Judge et al. [18] postulated that career success has intrinsic cues, like the person's own perception of success and self-satisfaction, and extrinsic ones, like awards, recognition or achievements. Since judgments about success in a creative profession like music are unavoidably subjective, we don't consider intrinsic factors and focus on objectively observable career accomplishments only.

Music industry criteria called "traditional markers of artist success" [12], like performance opportunities, labels, charts, awards, sales of recorded music or airplay, provide us with a number of possible directions for defining success of music artists. However, digitization has shaken these traditional markers—digital music has been linked to fall in record sales, airplay and charts no longer adequately measure popularity, given numerous streaming services and listenership outside of them—views on YouTube and/or illegal file-sharing. Given this, some researchers look at the popularity of music artists on digital delivery platforms like Last.fm, and formulate a forecasting problem to predict new song hits from the early adoption patterns of music listeners [33].

[2]The dataset is available at https://github.com/shushanarakelyan/forecasting_success

Success in post-digital music world can still be adequately represented by contracts with major labels. Music record labels are still important players in the industry—even though theoretically digital technologies allow artists to perform production, promotion and sales on their own, practically this doesn't happen very often [26]. Hence, in this work, forecasting success is operationalized as predicting the artists 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.

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 [10, 15], or modeling the evolution of a network [5, 19, 41]. 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 [23]. However, other authors [22] 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 [22, 23], or used techniques from recommender systems, such as low-rank matrix factorization and collaborative filtering [1, 6], and supervised learning approaches [4, 30]. 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* [16]—a modification to the *PageRank* [28] algorithm that tunes it towards bipartite graphs—to measure and predict the popularity of the artists and venues.

3 DATASET

Songkick[3] 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 music events.

Discogs[4] 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[5] 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]https://www.songkick.com/
[4]https://www.discogs.com/
[5]https://data.discogs.com/

3.1 Data Collection

Songkick does not provide a lookup directory of artists, nor there is a direct mechanism to get all 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. This can be either self-recorded or recorded under a contract with a music label. This strategy avoids introduction of 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. 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.

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),[6] or their subsidiaries, "successful", we provide a more detailed explanation for this choice in Section 5.1. 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.

Finally, we wanted to make sure that we have enough data about successful artists in the early stage of their career. Thus, in the last 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 with short-lived careers.

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

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

4 STATISTICS

In the following we provide some statistical analysis 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 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 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 a significant increase in the number of artists that have their earliest concerts recorded on Songkick in 2007 or later. From 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 one event, which, for the mostÂăpart, 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 in both number of concerts and number of venues is the UK. Figure 3 demonstrates distribution of venues and concerts in the 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 of all artists. As displayed in Table 1, n-grams of length 4 and 5 show some frequent routes of

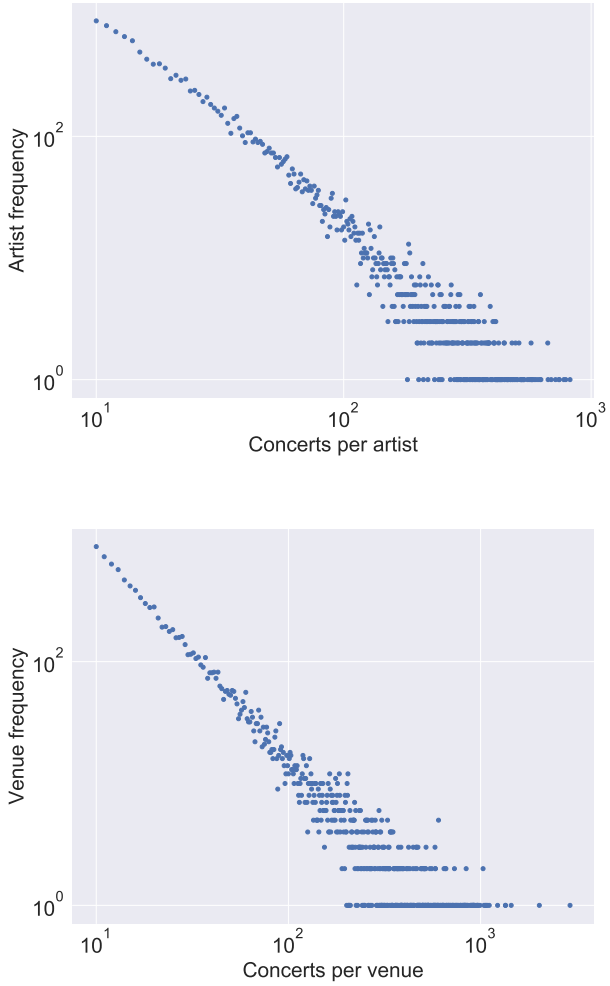

Figure 1: Heavy-tailed distributions of the number of concerts per artist (upper panel) and per venue (lower panel).

cities that artists take while touring. Following the distribution of the venues and concerts in the dataset, most frequent routes mostly include US cities. As demonstrated in Table 1, frequent routes 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.

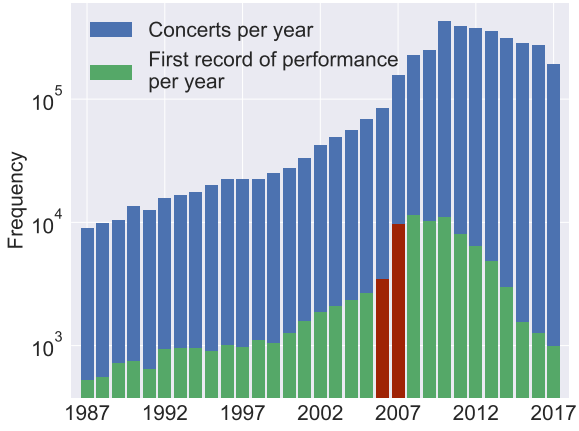

Figure 2: Total number of concerts per year in the dataset and number of artists that first appear on Songkick in a given year. The red bars illustrate the sudden change from 2006 to 2007 in the number of artists that first appeared on Songkick during those years. This can be explained by the fact that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

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 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

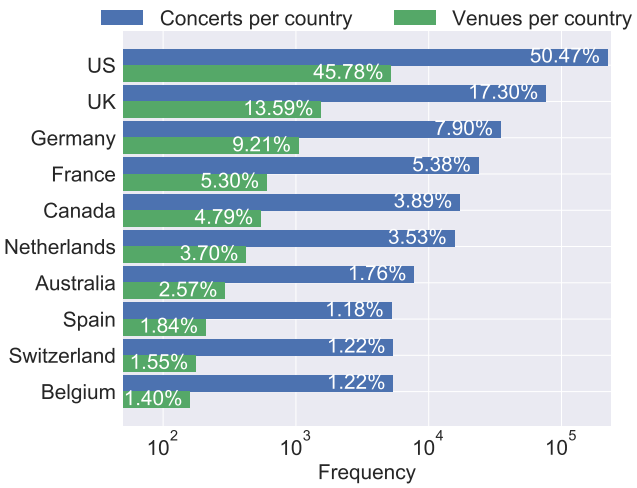

Figure 3: Log-scale distribution of concert frequencies in (i) the top 10 most active countries, and (ii) the number of distinct venues in those countries. A disproportionate preference toward English-speaking countries can be observed in the Songkick data, with United States, United Kingdom, and Australia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

to a release with a major label before the release itself happens. In the following discussion we refer to this task as *forecasting*.

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. The entries in the matrix are the numbers of times the artists performed at the venue. We used those vectors as features for the prediction and forecasting tasks.

In the forecasting task for any artist we did not include any concert that happened after the 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 number. It is worth noting that our labeling procedure yields a 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 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 can intuitively connect success of the artist to the number of their performances. We picked a baseline that would prove or disprove this scenario by using the number of concerts, scaled by the maximum number of concerts by an artist, as a proxy for probability for becoming successful.

Logistic Regression: As a base classifier in both prediction and forecasting experiments we used Logistic Regression from the scikit-learn library [31]. 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). Via cross-validation we discovered that best results are achieved when we use 750 components in prediction task and 1000 components in forecasting task.

For each model, we performed hyperparameter 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. Suggested baseline shows existing correlation between the number of concerts and prediction label, and this correlation is stronger in prediction task than in forecasting task. Next, simple logistic regression achieves 0.22 F1 score on the forecasting task and 0.4 on the prediction task. We can see that while reducing dimensions increase ROC AUC and F1 scores by several points in forecasting task, its improvement for prediction task is marginal.

The improvement in performance on the prediction task indicates there is a difference in distributions of artist performances before and after they record their first album with a major music label. This suggests the existence of change points in careers that are caused by recording with major labels, which corroborates our notion of artist's success. We expect that employing more sophisticated models for discovering change points 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.

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 (FCST/PRED LR+SVD)

Task	Model	P	R	F1	AUC
FCST	Baseline	0.07	0.26	0.11	0.60
FCST	LR	0.18	0.29	0.22	0.74
FCST	LR+SVD	0.18	0.35	0.23	0.78
PRED	Baseline	0.25	0.35	0.29	0.82
PRED	LR	0.36	0.45	0.40	0.86
PRED	LR+SVD	0.39	0.40	0.40	0.87

To see if we can explain part of those interactions, we formulate the artist-venue link prediction task. As in the forecasting artist success task, we consider here two configurations—*forecasting* and *prediction*. For this task we used the same affiliation network as in the previous task, but since we are interested in predicting new or hidden edges, we only use a binary affiliation matrix here.

In the previous task prediction experiments were performed to test whether or not our suggested definition of success is viable. For $(artist - venue)$ 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, or edges, 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 recursively removed all artists and venues that have less than 5 concerts associated with them in the training set. As a result we had 12,871 artists, 10,269 venues, 385,845 events in the training set and 43,122 events in the test set.

In the prediction task we kept the same set of artists and venues as described above for the forecasting task. We then 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. Results reported are averaged over such three random splits. We binarized all the links as we are only interested in predicting new links, i.e. new venues, where artist performs.

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 some popular heuristic scores for link prediction, a simple matrix factorization technique and node similarity based model, all described in the following.

Heuristic scores: Likelihood of a link existing between a pair of nodes is often approximated in terms of the number of common direct neighbors of that pair. However, a score calculated in this way will always be zero in a bipartite graph. Hence, we extended popular methods—Common Neighbors and Jaccard's coefficient—to use 2-hop neighbor sets of the pair instead of direct neighbors, as shown in Table 3, where $\mathcal{N}(u)$ is defined as the set of direct neighbors of node u, and $\hat{\mathcal{N}}(u) = \cup_{v \in \mathcal{N}(u)} \mathcal{N}(v)$ is the set of *neighbors of neighbors* of u. Another popular link prediction heuristic is Preferential attachment, which can be applied to a bipartite graph without any modifications.

Table 3: Heuristic scores for link prediction in bipartite graph for node pair (u, v). $\mathcal{N}(u)$ indicates direct neighbors of node u, $\hat{\mathcal{N}}(u)$ indicates neighbors of neighbors of u.

Common Neighbors ($CN(u, v)$)	$\lvert(\hat{\mathcal{N}}(u) \cap \mathcal{N}(v)) \cup (\hat{\mathcal{N}}(v) \cap \mathcal{N}(u))\rvert$
Jaccard's Coefficient	$\frac{CN(u,v)}{\lvert\hat{\mathcal{N}}(u) \cup \mathcal{N}(v) \cup \hat{\mathcal{N}}(v) \cup \mathcal{N}(u)\rvert}$
Preferential Attachment	$\lvert\mathcal{N}(u)\rvert \cdot \lvert\mathcal{N}(v)\rvert$

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 which venues are most relevant for the artist.

In this task, we used a simple yet popular collaborative filtering method based on matrix factorization—Singular Value Decomposition (SVD). To find the number of components for SVD, we used grid search—from 10 to 2000—and reported the result for 25.

Node similarity: Building and using graph representations is another approach that is often employed for link prediction. In our experiments we leveraged Deepwalk [32] for obtaining node representations and then used cosine similarity of a pair of nodes as an estimate for the probability of a link existing between them.[7]

Deepwalk is similar to training a Word2Vec model on a random walk sampled starting from every node in the graph. In our graph we gave preference to a larger number of short walks so we searched for the optimal number of walks of length 10. We report results for using 40 random walks. We then used cosine similarity of node representations as a proxy for probability of creating a new edge between those nodes.

Hyperparameters like number of hidden components in SVD and Deepwalk parameters in this task were only tuned for prediction task. We then used the same values for forecasting task. All parameters were estimated via grid search with 5-fold cross-validation, with 20% of all edges in each fold.

5.2.4 Task Summary. The results for the venue prediction task are presented in Table 4. As it can be seen, every method performs better on the prediction task than on forecasting, though for heuristic methods the improvement in performance is marginal. This hints that there might be a shift in artists' preferences for choosing a venue over time. It also indicates that while coarse statistics like Common Neighbors or Jaccard's coefficient are not affected much by those shifts, slightly more sensitive methods like SVD and node

[7] https://github.com/phanein/deepwalk

Table 4: Results for $(artist, venue)$ link prediction task, measured in Area Under Receiver Operating Characteristics curve (AUC).

Task	Model	AUC
FCST	Common Neighbors	0.87
FCST	Jaccard's coef	0.89
FCST	Preferential Attachment	0.79
FCST	SVD	0.81
FCST	Node similarity	0.84
PRED	Common Neighbors	0.91
PRED	Jaccard's coef	0.90
PRED	Preferential Attachment	0.84
PRED	SVD	0.91
PRED	Node similarity	0.90

similarity, that rely on the inner structure of the graph, are affected more. Yet, either that structure is not expressive, or the methods are not powerful enough, neither of those methods performs better than heuristic scores. Interestingly, four models out of five give performance of around 0.9 ROC AUC on prediction task. Out of all the methods we tried, Preferential Attachment has the lowest performance for both tasks.

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 [16]. This algorithm is a modification to the PageRank [28] 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.

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 [16].

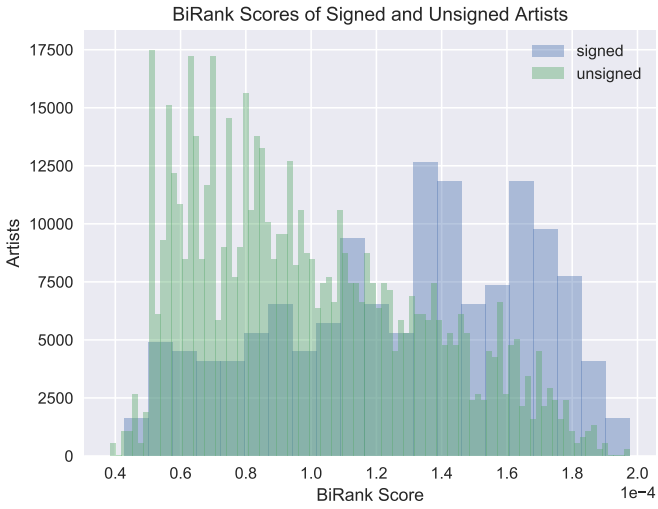

Figure 4: Histogram of signed and unsigned artists. Normalized to show relative frequency of BiRank scores.

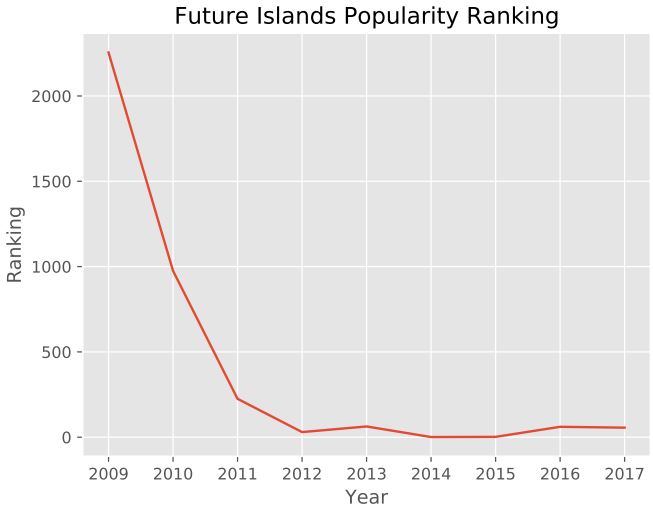

Figure 5: Trajectory of the group "Future Islands" through the lens of the BiRank score. The y-axis is the rank: lower is better. The BiRank score tracks the band's rise to popularity, culminating in the 2014 nomination of "breakthrough band of the year" by The Telegraph, suggesting that our framework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

The results of this experiment can be seen in Table 5. 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 artists 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,

Table 5: 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 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.

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."[8]

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.

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 would like to point out two potentially important limitations of the present work. First, the definition of success used here, while operationally useful, is by no means comprehensive. Indeed, many artists who work with independent labels, or specialize in commercially less-viable genres, can still have very successful and celebrated careers. And second, we note that despite its demonstrated usefulness, the dataset presented here is not perfect and is likely to 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 another important future task.

ACKNOWLEDGEMENTS

This research was supported in part by ARO (contract no. W911NF-12-R-0012) and 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.

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

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] Santa Agreste, Pasquale De Meo, Emilio Ferrara, Sebastiano Piccolo, and Alessandro Provetti. 2015. Analysis of a heterogeneous social network of humans and cultural objects. *IEEE Transactions on Systems, Man and Cybernetics: Systems* 45, 4 (2015), 559–570.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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*.

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

[10] 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.

[11] 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.

[12] M Evans et al. 2013. 'What Constitutes Artist Success in the Australian Music Industries?'. *International Journal of Music Business Research (IJMBR)* (2013).

[13] 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.

[14] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. 2018. Science of science. *Science* 359, 6379 (2018), eaao0185.

[15] 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.

[16] 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.

[17] Homa Hosseinmardi, Hsien-Te Kao, Kristina Lerman, and Emilio Ferrara. 2018. Discovering Hidden Structure in High Dimensional Human Behavioral Data via Tensor Factorization. In *HeteroNAM 2018: First International Workshop on Heterogeneous Networks Analysis and Mining*.

[18] Timothy A Judge, Chad A Higgins, Carl J Thoresen, and Murray R Barrick. 1999. The big five personality traits, general mental ability, and career success across the life span. *Personnel psychology* 52, 3 (1999), 621–652.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Rachel McLean, Paul G Oliver, and David W Wainwright. 2010. The myths of empowerment through information communication technologies: An exploration of the music industries and fan bases. *Management Decision* 48, 9 (2010), 1365–1377.

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

[32] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 701–710.

[33] 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.

[34] Anna Sapienza, Hao Peng, and Emilio Ferrara. 2017. Performance Dynamics and Success in Online Games. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. 902–909.

[35] 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.

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

[37] John Ternovski and Taha Yasseri. 2017. Social Complex Contagion in Music Listenership: A Natural Experiment with 1.3 Million Participants. *arXiv preprint arXiv:1711.05701* (2017).

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

[39] Burcu Yucesoy and Albert-László Barabási. 2016. Untangling performance from success. *EPJ Data Science* 5, 1 (2016), 17.

[40] Burcu Yucesoy, Xindi Wang, Junming Huang, and Albert-László Barabási. 2018. Success in books: a big data approach to bestsellers. *EPJ Data Science* 7, 1 (2018), 7.

[41] 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.

[42] 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.