The Kinomatics Australian Film Production Dataset: A curated dataset describing Australian feature film titles and personnel 1975-2022

P. Jones (University of Alberta): pete.jones@ualberta.ca
D. Verhoeven (University of Alberta): deb.verhoeven@ualberta.ca
S. Talalay Harvey (University of Alberta): smharvey@ualberta.ca
A. Kantoro (University of Alberta): kantorok@ualberta.ca
J. Martinez Garcia (University of Alberta): jcmarti1@ualberta.ca
V. Vo (University of Alberta): vv2@ualberta.ca

Abstract

This article presents a novel, extensive and thoroughly-documented dataset describing Australian feature films and the personnel filling ten key production roles on those films. The dataset is curated from public information in multiple sources and draws on further supplemental resources to verify, validate and consolidate this information. In total, the data describes 22,754 roles filled by 9,425 distinct people across 1,878 films, covering an important 47 year period in the Australian film industry. We outline how the dataset solves several problems for scholars interested in data that provides a historical record of the collaborative filmmaking process. In particular, to address concerns about known coverage problems with popular sources such as the Internet Movie Database, our data have undergone extensive manual checking to ensure that they are reliable as a source of information on a national film industry. Moreover, we have carefully and manually linked each person appearing in the dataset, which allows the dataset to provide a rich source of information for exploring the relationality of filmmaking collaborations. Our inclusion of ten key filmmaking roles further expands the utility of the dataset beyond existing datasets which tend to focus on actors and/or directors, writers and producers.

Keywords

Australian cinema; National cinema; Cultural production; Creative teams; Film production

Introduction

This paper introduces and describes a novel dataset containing information about several key crew roles involved in the production of Australian feature films from 1975 to 2022. Our goal in producing this dataset was to bring together information about who works on Australian feature films contained in multiple sources of digital and analogue records, and combine these into a comprehensive list covering the period of modern Australian cinema. Our efforts produced a dataset which describes more than 22,000 roles occupied by more than 9,400 unique people across almost 1,900 films. Despite its size, the dataset has been extensively checked and polished to ensure that it is reliable for scholars interested in researching feature film production in Australia. In this paper we detail the various methodological steps and decisions that we took in compiling this dataset, and provide a description of the basic characteristics and features of the dataset.

Problem

Film production is at its core a social process. Films are made by people who combine and sometimes recombine in project teams (Faulkner 1983; Faulkner and Anderson 1987; C. Jones and Walsh 1997). It is important for film studies to attend not only to gathering and archiving data on cultural texts, but also the people who were involved in their creation. We believe our dataset addresses three main challenges for accurately documenting the collaborative filmmaking process in the form of digital data.

First, our dataset goes beyond the Internet Movie Database (IMDb), a rich but limited data source, by prioritising other existing industry-led sources. IMDb is the most comprehensive single source describing films and their cast and crew, making it a valuable resource for academic researchers. However, the largely user-contributed information on IMDb has limitations and biases that structure and circumscribe the possibilities for research based on its data (Wasserman et al. 2015). Most notably there is a geo-political bias (films from the US feature most prominently), a temporal bias (more recent films have more metadata), and the data are impacted by the uneven frequency of missing and inconsistently reported data fields. Most significantly for our purposes is IMDB's limitations as a source of data for research on national industries outside of the United States (especially smaller film production settings such as Australia). IMDb is notably unreliable in relation to "country of origin" data which sometimes refers to the "country of production" and sometimes to the location of the film's production, post-production or investment companies and which never refers to the kinds of geo-political claims that would categorically position a film within a national or international (co-production) setting.. Because of these issues, we chose not to define a national cinema industry based on the "Country" field in IMDb. Instead, our data curation prioritised other industry-led data sources and compendia (as we detail in the next section); we used the information in IMDb primarily as an additional resource to assist our curation and disambiguation of those other sources, and as a primary source only when it provided information on roles that we did not already have from other sources.

Second, our dataset incorporates important crew roles that are rarely included in datasets which are used in research on film collaboration. Film production datasets typically focus on identifying who filled the so-called "key creative" roles of producer, director and writer, and/or the key acting roles (e.g. Ebbers and Wijnberg 2010; Liu and Ma 2022; Lutter 2015; Negro and Goodman 2015; Neuberger 2020; Pontikes, Negro, and Rao 2010; Rossman, Esparza, and Bonacich 2010). However, these roles provide a very incomplete picture of the relationships that underpin the film industry. In the case of actors, these roles are

typically distributed as the result of a casting process, and the actors themselves have very little control over this process; as such, it is not possible to infer collaborative intention from observing patterns of actor co-appearance. For key creatives, these three roles simply provide too little coverage of the overall film production process to give a meaningful window into creative teams. The extent to which screenwriters interact with producers and directors is variable and context-dependent, while producers are a famously difficult role to pin down in terms of overall involvement with the filmmaking process (Caldwell 2008; Cameron, Verhoeven, and Court 2010). In assembling the *Kinomatics Australian Film Production Dataset*, we wanted to ensure that we captured a wider range of roles in order to provide a more complete picture of the collaborations which structure the filmmaking process.

Third, our extensive manual review and cleaning of the data compiled from existing sources allows us to bring additional clarity and validity to what is often messy and speculative reporting of film projects. Typically, film researchers rely on production books and lists published in trade publications to document film productions. These sources are invaluable, but they also present challenges. In particular, they are often forward-looking, reporting on film productions that are early in development and still subject to change. These listed productions often evolve into other projects (which may also be documented, resulting in over-counting), and in many cases may not be completed or released at all. One of the goals of our dataset curation was to try to identify those productions which resulted in meaningful contributions to Australian cinema, in terms of both its release catalogue and its network of creative collaborations. This goal is reflected in our choice of methods, sources, and selection criteria, as we detail in the next section.

Methods

The first step in creating our dataset was to decide on the set of films that would constitute the corpus. We made the decision to only compile a dataset of live-action feature films. That is to say, we chose to exclude documentaries (unless they were significantly dramatised, and thus involved a similar set of crew roles to a typical fictional feature); to exclude animations (which generally involve a different set of roles and personnel); and to exclude short films and made-for-television films (which are generally not equivalent to feature films in terms of the industry personnel and practices involved in their production). In each of these cases, we felt that it would be more fruitful to focus on documenting a single industry (the feature film production industry) and its key personnel, and that adding in other parts of the screen sector would risk muddying the waters of the information we are primarily interested in. We also limited the dataset to films released between 1975 and 2022. We cut off the data period in 2022 as it was the latest full year for which we had complete information at the time of our data collection. The dataset begins in 1975, as this has been variously described by film historians as a "watershed" or "landmark" year for the Australian film industry in which a raft of significant events for Australian screen culture and institutions coincided (Goldsmith 2006; Verhoeven 2006).

Next, we determined how to define an "Australian film" for the purposes of this dataset. The question of what constitutes an Australian film is not straightforward to answer, and definitions differ according to time period and/or industry position (regulator, academic, archivist and so on). In film studies, since the revival of interest in the history of Australian cinema in the 1970s (which coincided with a revival in the industry itself) there have been several attempts to comprehensively define and capture Australian film industry production data. Many of these precede the era of digital databases (Murray 1994; Pike and

Cooper 1998; Verhoeven 1999), some were specifically developed as online resources in direct response to the limitations of IMDb coverage (Verhoeven n.d.). Each adopts a unique definition of what constitutes an "Australian" film and their coverage of Australian film production history is temporally limited and not updated, making analysis of longitudinal trends impractical if not impossible. To redress these limitations, we opted to use information maintained by Screen Australia (the Australian federal screen production funding agency) and served via its website in a list called the "*Screen Guide*" (Screen Australia n.d.) as the starting point for defining the set of films for inclusion in our dataset, though we supplemented this with information from other sources as we describe below.

For the *Screen Guide*, an Australian film is a project "under Australian creative control (i.e. where the key on-screen and/or off-screen elements are predominantly Australian and the project was originated and developed by Australians)" (ibid.). This definition includes projects under Australian creative control that "are 100 per cent foreign financed" as well as projects "where creative control is shared between Australian and foreign partners", with a balanced mix of "Australian and foreign elements in the key creative positions" (both official and unofficial 'co-productions'). We contacted Screen Australia to ask if they could provide the underlying data that is served via the *Screen Guide* web page in a format we could work with; they confirmed that this was not possible but we were permitted to gather the information from the web page ourselves. We then scraped the data into a tabular format, resulting in an initial comma-separated values (CSV) file containing each of the film-level fields described in the *Screen Guide* as columns and each role listed as part of the crew as a row. The nine roles found in the *Screen Guide*'s "crew" section are Producer, Executive Producer, Director, Writer, Editor, Cinematographer, Composer, Costume Designer and Production Designer, though not all of these roles are present for each film.

We merged into this dataset additional information contained in other datasets and sources of information on Australian cinema in the data period. The first of these additional sources is the *Production Book*, a directory providing extensive information about film and television productions and personnel for Australia (a version for New Zealand is also available). Listings are self-submitted by people in the industry and are reviewed and approved by the *Production Book* editors. In particular, we extracted information from the data table contained in the PDF published on the *Production Book* website (The Production Book, n.d.), which lists the title, production type, and year for each production from 1990 to 2019, as well as information on who filled certain roles (captured under the headings "Producers", "Script", "Director", "DOP" (Director of Photography), "Prod. Designer", "Costume Design", "Editor", and "Sound Design"). All but the last of these roles has an equivalent in the *Screen Guide* data, so we aligned these datasets by matching "Script" with "Writer", "DOP" with "Cinematographer", and we added "Sound Designer" as an additional role of interest, bringing the total number of roles in our dataset to ten. As with the *Screen Guide*, we filtered the data to only include films marked as "Feature" films, dropping series, telemovies and documentaries.

The next source we incorporated is the *Kinomatics Camera Departments (KCD) Dataset*, which was produced in association with the Australian Cinematographers Society for a separate research project (see Coate, Eltham, and Verhoeven 2023; Coles, Ferrer, and Zemaityte 2022; and P. Jones et al. 2024 for more details on this project and the data). Although the *KCD Dataset* is primarily concerned with camera department roles (headed by the cinematographer), it also contains information about writers, directors and producers. The *KCD Dataset* was used to add any additional information about who filled our ten key

roles that was not captured in either the *Screen Guide* or *Production Book* datasets. During the process of merging these sources into a single common data table, IMDb IDs were manually added to each film where we could find one, enabling future researchers using this dataset to merge in additional information from IMDb (or other data sources which make use of the IMDb ID as a key) if desired.

In order to ensure that our dataset was not missing any important titles which one might expect to be in a dataset of Australian feature film production from this period, we compiled a list of all films which were nominated for a major award (see the technical documentation for details) by the Australian Film Institute (AFI) Awards (later renamed the Australian Academy of Cinema and Television Arts (AACTA) Awards) from 1975 to 2022. Of the 492 films nominated during this period, 482 (95 %) were already in our dataset. The remaining 17 film titles were reviewed for eligibility, yielding a total of 11 films flagged for inclusion. In addition, we used two film history books which are known as Australian film "bibles" for their authoritative curation of Australian feature films. The first of these books is Pike and Cooper's *Australian Film 1900-1977* which provides comprehensive coverage of Australian film production including crew details and plot synopses (Pike and Cooper 1998). The second is Scott Murray's *Australian Cinema 1978-1994* which picks up where Pike and Cooper left off (Murray 1994). For the 11 award-nominated films and the 35 films gleaned from the Murray and Pike and Cooper books, we pulled information from IMDb for the ten key roles and appended this to our data.

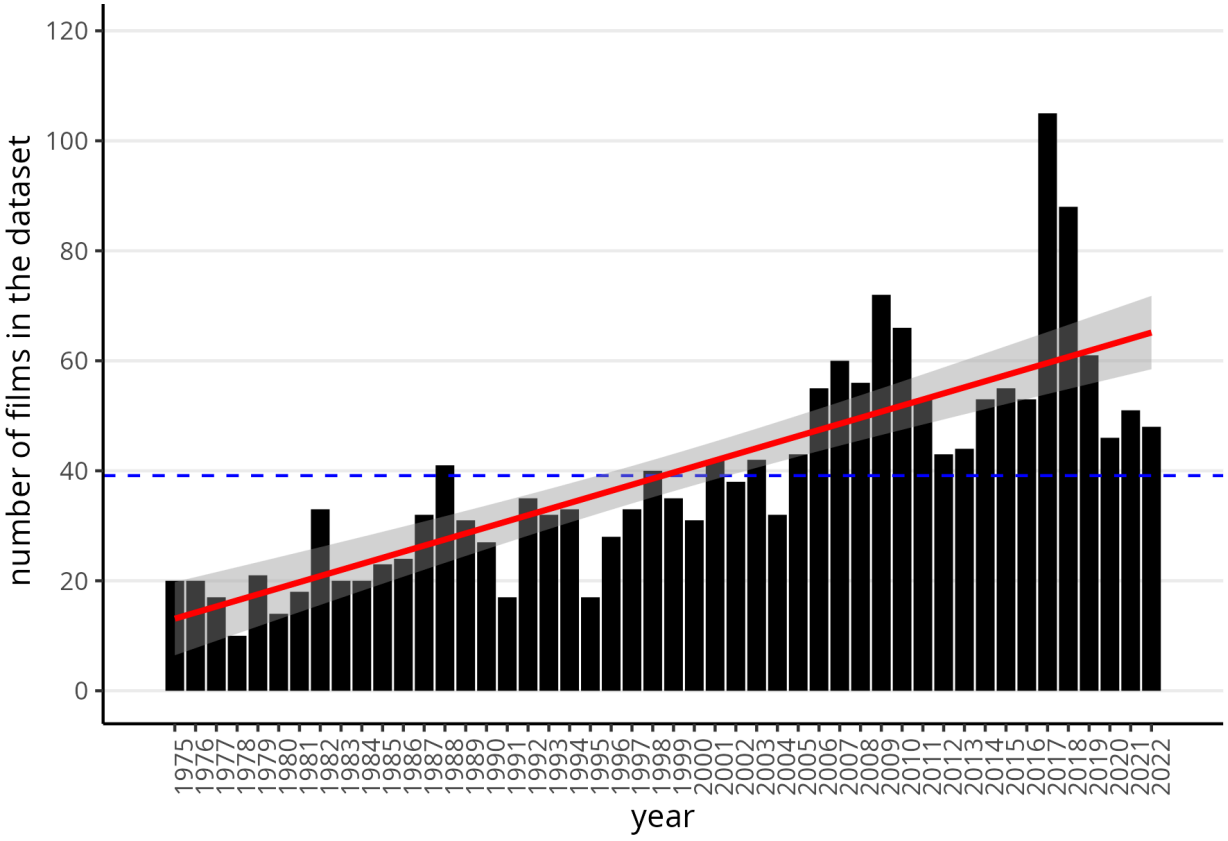
At this stage, we manually checked, cleaned and validated the merged dataset. We provide more detail on our data cleaning and verification process in the dataset's technical documentation but, briefly, this involved interoperating the data in a common format, removing duplicated information, and inspecting and resolving any inconsistencies in the resulting data. Here, we closely consulted the various aforementioned data sources as well as the website Ozmovies.com (which contains useful digitised archives of the films' credits as they appear in the video) and a further authoritative edited volume *Twin Peeks* (Verhoeven 1999) which compiles production details for Australian feature films (as well as those produced in New Zealand) up to the year 2000. The most intensive aspect of our manual data validation involved connecting people across films using persistent identifiers - a particularly tricky task given the inherent "slipperiness" of names that makes them problematic as a basis for identification in cultural data (Cutter, Fensham, and Sumner 2023). We then filtered out any data which, following validation and cleaning, were discovered to fall foul of our inclusion criteria for the dataset. For the remaining films, we used IMDb to pull in any missing information on who filled the ten key roles; for the people identified from this process, we cleaned, verified and integrated the data following the same procedures we used for the main data sources. Following the verification and integration of this last tranche of data, we arrived at the final dataset, which we describe in the next section.

Data

The dataset and its documentation are deposited in Zenodo, and shared under the CC BY-NC-SA 4.0 license. The dataset is primarily presented through two CSV files: "roles.csv" (wherein each row corresponds to an instance of a person filling a role on a given film, and information about the role, including the identifiers for the person and the film, are captured in the columns), and "films.csv" (in which each row corresponds to a film, and the columns correspond to film-level variables). Detailed tables breaking down each of these data files variable-by-variable can be found in the technical documentation in the data repository.

In total, the dataset contains information on 22,754 occupied roles, wherein 9,425 unique people work on 1,878 unique films. The mean number of people attached to each film (excluding three anthology films, which skew this figure) is 9.83 (standard deviation = 4.76), while the median is 9. The mean number of films worked on per person is 1.98 (standard deviation = 2.8), and 71% of people only work on one film in the dataset. Two features of the dataset should be taken into consideration by any users of the data. First, the number of films in the dataset is not consistent over time, and there is a notable upward trend in the number of films per year (see Fig. 1).

Fig. 1. Number of films in the dataset for each release year 1975-2022. Dashed blue line indicates the overall mean; solid red line indicates a simple linear regression of number of films on year.

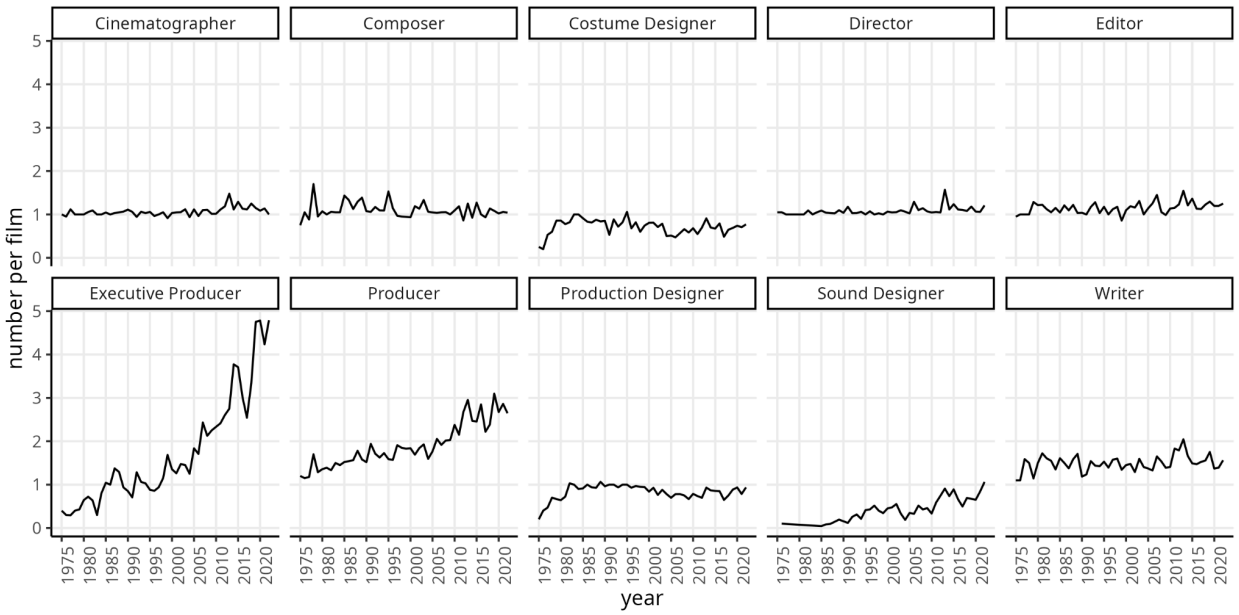


Second, the number of people identified in each of the ten roles varies. Table 1 lists the total number of times each role appears in the dataset. Overall, production roles are the most frequent, while sound and production design roles are the least common. Furthermore, the relative frequencies of roles also vary over time: there is a notable increase in the number of executive producers identified per film in the last couple of decades, though many other roles remain constant (see Fig. 2).

Table 1. The total number of times each role appears in the data. The “number per film” value is the total number divided by the overall number of films (1,878).

Role	**Total number**	**Number per film**
Executive Producer	4042	2.15
Producer	3871	2.06
Writer	2839	1.51
Editor	2193	1.17
Director	2054	1.09
Composer	2037	1.08
Cinematographer	2036	1.08
Production Designer	1550	0.83
Costume Designer	1304	0.7
Sound Designer	828	0.44

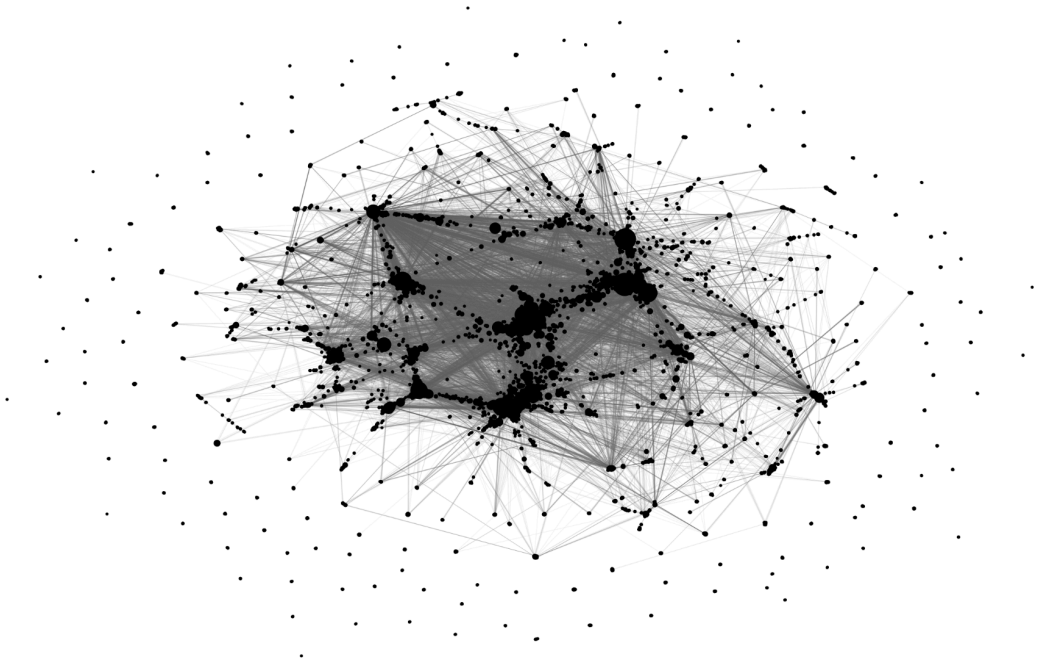
Fig. 2. Number of times each role is filled on a project over time.



The key variables for scholars interested in constructing collaboration networks from the data are the “person_id” and “film_id” columns from the “roles.csv” table. These two variables allow for the construction of a network representation of the data wherein people can be connected based on the films

that they worked on together. To illustrate this relational data structure, we present a visualisation of the full collaboration network in Fig. 3.

Fig. 3. Network visualisation for the one-mode projection of the person-to-film affiliation matrix for all films in the dataset. Nodes represent people, edges between the nodes represent working on the same project. Nodes are sized proportionally to the total number of films the person worked on, and edge opacity is scaled to edge weights



Conclusion

This paper has introduced the *Kinomatics Australian Film Production Dataset*, a curated list of Australian feature film productions and the people who filled ten key creative roles on them. The dataset has obvious utility for researchers such as ourselves who are interested in studying historical patterns of collaboration among filmmakers. However, the dataset also presents opportunities for scholars to approach the history of production in Australian cinema from multiple angles. While scholars interested in Australian film production specifically will most clearly benefit from the dataset, we believe it provides a rich source of information for anyone interested in national cinema or film production dynamics in general. Moreover, the inclusion of IMDb IDs (which are found in many film-related data sources, including but not limited to those derived from IMDb itself) enables scholars to interoperate the dataset with other datasets, expanding the range of research questions the dataset can be used to explore. While we strived to be as comprehensive and diligent as we could in our curation of the data, it is inevitable that a dataset as large as this which describes a phenomenon as messy and diverse as filmmaking will contain some errors and

oversights. We invite readers that find any such errors to contact either of the dataset's two lead authors so that we can update the dataset where necessary.

References

Caldwell, John Thornton. 2008. *Production Culture: Industrial Reflexivity and Critical Practice in Film and Television*. Duke University Press. https://doi.org/10.1215/9780822388968.

Cameron, Allan, Deb Verhoeven, and David Court. 2010. 'Above the Bottom Line: Understanding Australian Screen Content Producers'. *Media International Australia* 136 (1): 90–102. https://doi.org/10.1177/1329878X1013600112.

Coate, Bronwyn, Ben Eltham, and Deb Verhoeven. 2023. 'By a Long Shot: Addressing Power, Devaluation and Discrimination in a Toxic Cultural Workforce'. *Australian Journal of Management*. https://doi.org/10.1177/0312896223117937.

Coles, Amanda, Justine Ferrer, and Vejune Zemaityte. 2022. 'A Wider Lens: Australian Camera Workforce Development and Diversity'. https://cinematographer.org.au/wp-content/uploads/2022/07/A-Wider-Lens-report-final.pdf.

Cutter, Nat, Rachel Fensham, and Tyne Daile Sumner. 2023. 'The Slipperiness of Name: Biography and Gender in Australian Cultural Databases'. *Gender & History* n/a (n/a). https://doi.org/10.1111/1468-0424.12699.

Ebbers, Joris J., and Nachoem M. Wijnberg. 2010. 'Disentangling the Effects of Reputation and Network Position on the Evolution of Alliance Networks'. *Strategic Organization* 8 (3): 255–75. https://doi.org/10.1177/1476127010381102.

Faulkner, Robert R. 1983. *Music on Demand: Composers and Careers in the Hollywood Film Industry*. New Brunswick: Transaction Books.

Faulkner, Robert R., and Andy B. Anderson. 1987. 'Short-Term Projects and Emergent Careers: Evidence from Hollywood'. *American Journal of Sociology* 92 (4): 879–909. https://doi.org/10.1086/228586.

Goldsmith, Ben. 2006. 'A Bureaucratic Cinema: The Fraught Relationship between Government and Film'. *Metro*, 2006.

Jones, Candace, and Kate Walsh. 1997. 'Boundaryless Careers in the US Film Industry: Understanding Labor Market Dynamics of Network Organizations'. *Industrielle Beziehungen* 4.

Jones, Pete, Deb Verhoeven, Aresh Dadlani, and Vejune Zemaityte. 2024. 'She Must Be Seeing Things! Gender Disparity in Camera Department Networks'. *Social Networks* 76 (January): 120–34. https://doi.org/10.1016/j.socnet.2023.09.004.

Liu, Yixuan, and Yifang Ma. 2022. 'Quantifying Award Network and Career Development in the Movie Industry'. *Frontiers in Physics* 10. https://www.frontiersin.org/articles/10.3389/fphy.2022.902890.

Lutter, Mark. 2015. 'Do Women Suffer from Network Closure? The Moderating Effect of Social Capital on Gender Inequality in a Project-Based Labor Market, 1929 to 2010'. *American Sociological Review* 80 (2): 329–58. https://doi.org/10.1177/0003122414568788.

Murray, Scott, ed. 1994. *Australian Cinema*. St. Leonards, NSW, Australia: Allen & Unwin in association with Australian Film Commission.

Negro, Giacomo, and Sasha Goodman. 2015. 'Niche Overlap and Discrediting Acts: An Empirical Analysis of Informing in Hollywood'. *Sociological Science* 2: 308–28. https://doi.org/10.15195/v2.a15.

Neuberger, Joan. 2020. 'Centrality and Centralisation: A Social Network Analysis of the Early Soviet Film Industry, 1918-1953'. *Apparatus. Film, Media and Digital Cultures of Central and Eastern Europe*, no. 10 (October). https://doi.org/10.17892/app.2020.00010.177.

Pike, Andrew, and Ross Cooper. 1998. *Australian Film 1900-1977: A Guide to Feature Film Production*. 2nd ed. Oxford: Oxford University Press.

Pontikes, Elizabeth, Giacomo Negro, and Hayagreeva Rao. 2010. 'Stained Red: A Study of Stigma by

Association to Blacklisted Artists during the “Red Scare” in Hollywood, 1945 to 1960’. *American Sociological Review* 75 (3): 456–78. https://doi.org/10.1177/0003122410368929.

Rossman, Gabriel, Nicole Esparza, and Phillip Bonacich. 2010. ‘I’d Like to Thank the Academy, Team Spillovers, and Network Centrality’. *American Sociological Review* 75 (1): 31–51. https://doi.org/10.1177/0003122409359164.

Screen Australia. n.d. ‘What Is The Screen Guide?’ The Screen Guide - Screen Australia. n.d. https://www.screenaustralia.gov.au/the-screen-guide/what-is-the-screen-guide.

The Production Book. n.d. ‘Australian Film Productions, 1990-2019’. The Production Book. Accessed 24 April 2024. http://www.productionbook.com.au/.

Verhoeven, Deb, ed. 1999. *Twin Peeks: Australian and New Zealand Feature Films*. Melbourne, VIC: Damned Publishing.

———. 2006. ‘1975: The Unease of Passing Milestones’. *Metro*, 2006.

———. n.d. ‘Bonza - National Cinema & Television Database’. n.d. https://bonzadb.com.au/.

Wasserman, Max, Satyam Mukherjee, Konner Scott, Xiao Han T. Zeng, Filippo Radicchi, and Luís A. N. Amaral. 2015. ‘Correlations between User Voting Data, Budget, and Box Office for Films in the Internet Movie Database’. *Journal of the Association for Information Science and Technology* 66 (4): 858–68. https://doi.org/10.1002/asi.23213.