# Imitation vs. Innovation: Product Similarity Network in the Motion Picture Industry

Yanhao Wei*
University of Pennsylvania

This draft: September 8, 2015
For the latest draft, visit:
https://sites.google.com/site/yanhaomaxwei/research.

**Abstract** This paper studies product entry when firms learn about a potential product from the market performance of previous similar products. Focusing on the U.S. motion picture industry, we construct a network capturing the similarity amongst the movies released in the last decades. We develop a model of how the network evolves. Risk-averse firms make investment choices on candidate products that arrive over time and can be either novel or similar to various previous products. By estimating the model and conducting counter-factual experiments, we demonstrate that learning matters and provide insights on the innovation vs. imitation tradeoff. In particular, we find that one firm benefits substantially from the learning of the other firms. We find that big-budget movies benefit more from imitation, but small-budget movies favor novelty. This leads to interesting market dynamics that cannot be produced by a model without learning.

*Email: yanhao@sas.upenn.edu. I am indebted to the members of my doctoral committee, Holger Sieg, Eric Bradlow, Joseph Harrington, Katja Seim and Christophe Van den Bulte, for their support on this research. For their helpful suggestions, and also not blaming them for any mistakes in the paper, I want to thank Bryan Bollinger, Ron Berman, Jehoshua Eliashberg, Peter Fader, Hanming Fang, Xiang Fang, Jian Ni, Devin Reilly, Hongxun Ruan, Jagmohan Raju, James Roberts, Andrew Sweeting, Francisco Silva, Qiaowei Shen, Petra Todd, Rakesh Vohra, Pinar Yildirim and Weilong Zhang. I also would like to thank the participants at the 37th Marketing Science Conference and various seminars at the Department of Economics and the Wharton School of the University of Pennsylvania for their comments.

# 1 Introduction

In many industries, new products keep rolling out at a fairly frequent pace, so managers need to constantly decide what products to introduce and anticipate market reception: Should I introduce a novel product or imitate some existing products? Is a particular potential product going to be received well? Examples of such industries include motion pictures, book publishing, video games, TV shows, software development, cell phone manufacturing, apparel, and even scientific research. In these examples, much can be learned from the market performance of past similar products in regards to whether a new product will be successful. So while firms decide what products to introduce, these products in turn affect the product-line decisions of the firms.

This paper focuses on the U.S. motion picture industry to study firm learning from previous products. The movie industry is a popular setting for marketing and economic research. It is also a place where product similarity plays a salient role. Characterized by a high degree of uncertainty, the industry sees a wide range of returns on investment. For instance, films like E.T. The Extra-Terrestrial grossed $360m domestically on an $11m budget, whereas The Golden Compass lost $110m from a $180m budget.[1] Industry experts observe that such an environment offers "no magic formula for a commercial movie," and the only viable strategy seems to be "emulating prior successes" (Squire (2005), p.4).[2]

To better understand how learning gives rise to similarity among products, we develop a model of product entry that focuses on firm-side uncertainty and learning about product quality. A product's market reception is determined by its "latent quality" as well as its observed characteristics such as budget size, genre and star power. The latent quality captures the consumer preferences on characteristics such as visual effects, acting, theme, storyline, narrative method, pace and music style. When a production company decides whether to green light a potential movie, typically over one year before the release date, it is uncertain about the movie's latent quality. However, the latent quality is signaled by the market performance of the similar movies that have been released. At any time, each firm holds a portfolio of in-production movies; they are risk averse and seek to maximize the risk-adjusted total profits from the future releases of the movies in the portfolio.

Because the full characteristic space is generally high-dimensional, it is a challenge to capture the similarity amongst products in an empirically tractable way. We take a network approach and use a similarity network where a link between two products represents proximity in their product characteristics. To model the supply of potential products, we apply ideas from the literature on evolving random networks (Newman (2003), Jackson (2010)) and specify a stochastic process where potential movies continuously arrive and link, or "attach," to the existing network. Observed characteristics and latent quality are determined conditional on the attachment. The

[1] On movie industry uncertainty, see also De Vany and Walls (1996, 1999).

[2] See also "Hollywood learns originality does not pay." May 29, 2015, *Financial Times*.

process provides the set of candidate products from which firms can select, and its stochastic nature means that the candidate can be similar to few or many existing movies, offering opportunities of both innovation and imitation.

We bring the model to data. To construct the similarity network, a head-on approach calls for calculating the proximity between movies in terms of their characteristics, many of which are difficult to observe. We take an indirect approach and use consumer data, with the idea that consumer preferences are correlated between similar movies. Specifically, we explore the "consumers who liked this also liked" feature on IMDb.com and a similar feature on Amazon.com to construct a network among the movies released in the U.S. in the last decades.[3] Through reduced-form analysis, we find that previous similar movies are much more predictive of a movie's market performance than the covariates commonly used in movie studies (e.g., budget, genre, star power). We also find evidence that suggests firm-side learning and risk aversion.

The paper proceeds to estimate the model with the method of simulated moments and conduct counter-factual experiments. Several insights are derived. First, we demonstrate that learning matters for the movie business. For the movies in the data, learning reduces the variance in the latent quality at the time of green-light decisions by over 60 percent, on average. Learning allows a firm to produce big-budget movies and maintain a high level of latent quality on its movies. Learning also has indirect benefits to other firms, as they are capable of imitating the products of each other. We find that for a major studio, the indirect benefits from the learning of the other firms are comparable to the direct benefits from its own learning.

Other insights pertain to the balance of innovation versus imitation. We find that whether to imitate or innovate crucially depends on the investment size. Big-budget movies benefit more from imitation as a way to reduce risks. However, small-budgets favor novelty because risk is of less concern and a higher level of uncertainty increases the chance for them to make a big hit. In a related counter-factual, we find that a mere increase in the arrival rate increases not only the number of movies produced per year, but also the average budget size of these movies. The reason is that the larger number of movies provides more room for imitation. In another counter-factual, we find that a higher level of risk aversion can actually increase the level of innovation. One cause is that the firms shift to smaller-budget movies where innovation is favored. These results provide some unique insights to the rise of blockbusters and the debate surrounding it.[4]

In terms of general insights, this paper adds to the studies on product networks. Compared with the widespread attention on social networks, it is perhaps surprising that there are only a handful papers on product networks, including Dellarocas et al. (2010) on news reports and Oestreicher-Singer et al. (2013) on online recommendation.[5] The paper also adds to the studies

[3] Oestreicher-Singer and Sundararajan (2012) also study the co-purchase network on Amazon.com, but focus on how the visibility of the network can alter demand spillovers across the linked products.

[4] See "Are Blockbusters Destroying the Movies?" *New York Times*, Jan. 6, 2015.

[5] Some network papers study products that are strongly associated with people, such as user-generated contents. See Mayzlin and Yoganarasimhan (2012), Lu et al. (2013) and Shriver et al. (2013).

on firm learning about demand. Hitsch (2006) and Shen and Liu (2014) model how firms learn from product sales after its launch and exit optimally. Toivanen and Waterson (2005), Shen and Xiao (2014) and Yang (2014) examine how firms learn from the *choices* of each other in the context of market entry and exit in fast food chains. Aside from the differences in learning channels and contexts, these studies do not consider the notions of similarity, risk aversion and imitation-innovation trade-off, which are the focus of this paper. More broadly, the paper is related to the literature on learning models (Ching et al. (2013)).

In terms of the application, the paper contributes to the literature on the motion picture industry, a domain that has high economic importance and public interest. Related works include Ravid (1999) which studies the effects of star power on box-office revenue, Eliashberg et al. (2007) which applies textual analysis to investigate what types of storyline are more likely to produce successful movies, and Goettler and Leslie (2005) which documents the risk aversion of studios. Overall, extensive modeling work has been done on box-office demand, theatrical release and exhibition,[6] leaving much less attention on the green-light decisions. To the best of our knowledge, this paper provides the first empirical model for product-line decisions in motion pictures.

The rest of the paper is organized as follows. Section 2 describes the data. Section 3 presents a reduced-form analysis. Section 4 develops the model. Section 5 describes how we estimate the model. Section 6 presents the model estimates. Section 7 conducts counter-factual experiments. The last section concludes and provides guidance for further research.

## 2 Data

**Data Sources** We mainly use two categories of data. The first is on movie characteristics. These include those commonly used in studies of motion pictures: title, release date, language, region, genre, MPAA rating, production budget, writers, directors, leading actors, and domestic box-office revenue. Domestic box-office revenues only account for a part of a movie's total revenues. However, it heavily influences revenues in subsequent markets, and is widely used in measuring the market performance of movies (Eliashberg et al. (2006), Einav (2007)).[7] Because we want to study firm-side decisions, we also collect data on the production companies and production start date of each movie.

Most of the movie characteristics are collected from the Internet Movie Database (IMDb.com). Additional data on box-office revenues are collected from Boxofficemojo.com, which offers better separation between the revenues from multiple releases if the movie was ever re-released. In this paper we focus on the box office at the first release. For a small number of movies whose budget sizes are missing on IMDb.com.com, we are able to collect them from Wikipedia.com.

[6] For a comprehensive survey, see Eliashberg et al. (2006).

[7] We collected data on international box-office revenues (roughly half the coverage). A polynomial regression against the domestic box-office revenues shows a mostly linear relation.

To account for inflation, we collect data on yearly price level (Consumer Price Index) from the U.S. Bureau of Labor Statistics. To calculate market share from box-office revenues, we collect yearly data on average theater ticket price from The-numbers.com and U.S. population from the Census Bureau.

Production start date is unavailable for roughly one third of the movies. So we estimate a relation between the budget size and production period (time elapsed from start date to release date) and use it to impute the start date. On average a movie takes slightly more than one year to produce. The estimated relation has an U shape, where the production periods of medium-budget movies are the shortest. We have experimented with a few other ways of imputation and found very little changes to the results.[8]

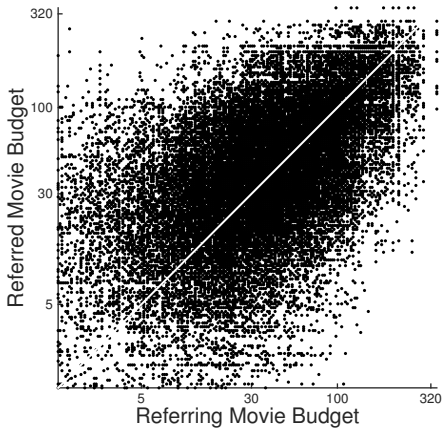
The second category includes the data that allow us to construct a similarity network amongst these movies. While a head-on approach calls for calculating the proximity between movies in their full characteristic space, this requires very rich data that are hardly available. Another approach is using revealed preferences. Researchers have exploited panel data on consumer purchases to uncover product positions in a latent space (Chintagunta (1994), Elrod and Keane (1995), Goettler and Shachar (2001)). We will not use a detailed model of consumer choice of multiple brands as in these studies. Instead, we simply construct the similarity network directly from co-purchase data.

Specifically, we make use of the "people who liked this also liked" feature on IMDb.com. Through this feature, the webpage for each movie refers viewers to the pages of up to 12 other movies. The video-on-demand service on Amazon.com offers a similar feature under the title "customers who watched this also watched," where each movie refers to up to 20 other movies.[9] We combine the two data sources and define a link between two movies whenever one refers to the other on either website. We collect the reference data with a web crawler. We visit the websites without logging in any account and disable the browser cookie to prevent the references from being tailored for a particular browsing history.

Websites usually closely guard the details of their algorithms for generating recommendations. IMDb.com states that their formula uses factors such as user votes, genre, title, and keywords. Amazon.com uses item-to-item collaborative filtering that builds similar-item table by finding items that customers tend to purchase together (Linden et al. (2003)). Given these high-level descriptions, the reference data seems suitable for the purpose of constructing a similarity network. To further check the validity of our similarity measure, we provide a preliminary examination of the network after discussing the sample selection, and more analysis in Section 3. We have also run the analyses throughout the paper with the network constructed from the Amazon data only, but have not found qualitative differences in the results.

[8] We have experimented with (i) estimating production period as a function of budget, genre, rating and crew power, (ii) setting a flat production period, e.g., 1.25 years, for all the movies.

[9] Not every movie is available at the video-on-demand service on Amazon.com. However, for our sample, only a small proportion (less than 5%) is not covered. The data were collected in March, 2014.




A dot with location $(x, y)$ represents a recommendation from a movie with budget $x$ to another movie with budget $y$ on Amazon.com or IMDb.com. Budget sizes are normalized by CPI to be in 2014 million dollars. The axis scale is nonlinear. The entire sample (1975-2012) is included.

Figure 1: Budgets of the Referred Movies against Referring Movies

**Sample Selection** We focus on the movies released in the U.S. that started production between 1995 and 2012 (included). The release dates of these movies extend to 2014. We exclude the "micro-budget" movies, which are defined as those with a budget less than 1 million in 2014 dollars. The mechanism behind the production and distribution of the "micro-budget" movies is likely quite different from that of the bigger movies. We have to leave out the movies for which either budget or domestic box-office gross is unavailable. Such movies are typically the ones without significant theatrical release in the U.S. In the end, we have a sample of 3,036 movies.

It is a good idea to include older movies as the "initial state" for our analyses. This is particularly important for the movies that started closely after 1995 because otherwise they would have no previous similar counterparts and appear all original. Movies that came later in the sample period are much less likely to be similar to these earlier movies. We are able to include 1,354 movies from 1975 to 1994 as the initial state. The small sample size is partly due to the fact that fewer movies per year were produced at that time, and partly due to a significant drop in data coverage as we go before the early 1980s. We have also tried using 1985-1994 as the initial period, but have not found qualitative changes to the results.

**The Network** Recall that a link between two movies is defined by a reference from either movie to the other on IMDb or Amazon. There may be concern that the recommendations are heavily biased in favor of "big" movies, or blockbusters, either because these movies are watched by almost everyone. However, recommendation algorithms usually compensate for the popularity of each item (Linden et al. (2003)). Figure 1 plots the budgets of the referred movies against those of the referring movies. A dot represents a reommendation from one movie to another. The budget size of the first movie is given by the horizontal position of the dot, while the budget

Table 1: Movie Pair Characteristics and Links

| | All Dyads | Linked Dyads | Logit Model |
|---|---|---|---|
| Intercept | | | -5.72 (.035) |
| Same Production Company | 3.60% | 10.6% | 0.650 (.048) |
| Same Rating | 35.3% | 59.7% | 0.910 (.028) |
| Same Genre | 19.4% | 51.0% | 1.26 (.028) |
| Same Leading Actor(s) | 0.603% | 27.8% | 3.85 (.034) |
| Same Director(s) | 0.099% | 8.77% | 3.53 (.09) |
| Same Writer(s) | 0.081% | 7.26% | 2.51 (.09) |
| Difference in Release Time | 9.76 | 5.37 | -0.113 (.003) |
| Difference in Log Budget | 1.22 | 0.762 | -0.532 (.020) |
| Pseudo-$R^2$ | | | 0.244 |
| $N$ | 9.63e7 | 2.91e4 | 9.63e7 |

The unit of release time is year. Budget is in 2014 million dollars. In case that there are multiple production companies for one movie, we use the first-listed one. The same applies to genre. The last column is a logistic regression using pair characteristics to predict linkage. Pseudo-$R^2$ equals 1 minus the ratio between residual deviance and null deviance. The entire sample (1975-2012) is included.

of the second movie is given by the vertial position. If recommendations are biased towards big movies, most of the dots should be above the 45° line. However, the plot is nearly symmetric, indicating that there is limited, if any, bias.

There is also concern that co-watching data is ex post information which studios do not possess when they green-light movies. With respect to this, notice that our purpose of using the co-watching data is to back out the similarity amongst movies, which we assume that the firms always understand. In fact, the network that we construct should be transparent enough for an experienced studio executive or movie producer to figure out. For example, the famous WWII movie *Saving Private Ryan* links to *Schindler's List* by the same director, and Vietnam War movies *We Were Soldiers* and *Full Metal Jacket*. A bit more sophisticatedly, it also links to the *The Patriot*, a movie on American Revolution but by the same writer, and *The Shawshank Redemption*.[10] However, it does not link to, for example, *The English Patient* or *The Reader*, which also use WWII as background but lean toward a more romantic theme.

Table 1 provides some descriptive statistics of the network. We see that a linkage clearly indicates proximity of the two nodes in terms of their observed characteristics. For example, among all the possible pairs of movies, 19.4% belong a same genre; the percentage nearly triples when it is among all the linked pairs. The last column presents a logit model that uses pair characteristics to predict linkages. All the coefficients are statistically significant. However, the Pseudo-$R^2$ of

[10]Both *Saving Private Ryan* and *The Shawshank Redemption* belong to the top guy-cry movies selected by *Entertainment Weekly*, 2005.

the model is modest, suggesting that the network contains rich information on other dimensions than these observed characteristics, a point that we will corroborate in the reduced-form analysis.

## 3 Reduced-form Analysis

In this section we present some model-free results. In particular, we examine to what extent the performances of similar movies are correlated. Market performance is measured here by ROI (return of investment, defined as the ratio between domestic box-office revenue and budget). In particular, we are interested in seeing if the ROIs of the previous similar products are able to predict the ROI of a product in addition to its observed characteristics. We also explore what movies are more likely to be imitated by or to imitate others. Aside from being interesting in their own right, the results also motivate the model in Section 4.

Table 2 examines the possible predictors of a movie's box-office performance. Specifically, Column 1 regresses the log ROI on a time trend, genre, MPAA rating, crew power and log budget.[11] These covariates are common in the studies of the industry. Notice that there are no significant effects of the "star power," which is consistent with the finding in Ravid (1999) that stars capture their expected economic rent. Notice that the $R^2$ of the regression model is very low, indicating that the market performance is hardly explained by the observed characteristics. It is worthwhile to point out that though the budget size hardly explains the ROI, it explains substantial variation in the box-office revenue. The $R^2$ rises to 0.54 if we use the log box-office revenue as the dependent variable, which is comparable to the $R^2$'s found in previous studies.[12]

In Column 2, we add a spatial lag term which equals the average log ROI of the "precursors." A precursor for movie $j$ is a movie that is similar to $j$ and precedes $j$ in terms of released date. The coefficient estimate of the lag term is positive and significant, and the $R^2$ is greatly improved when compared with Column 1. This indicates that the network captures the proximity amongst movies along the unobserved characteristics that are capable of affecting the ROI. In particular, notice that coefficient on writer becomes much smaller and insignificant. A writer is likely to carry her style of storytelling from one movie to another, which seems to have been picked up by our similarity measure.

In Column 3, we drop genre, rating and crew as covariates but keep the spatial lag. The decrease in $R^2$ is only marginal. This indicates that the ROIs of the previous similar movies are better predictors of the ROI of a movie than these observed characteristics. Also notice that the coefficient of the spatial lag is slightly increased. This indicates that our similarity measure has incorporated, to certain extent, the proximity amongst movies with respect to the dropped

[11] Some may be concerned with the fact that budget appears on both sides of the regression. An alternative regression where the dependent variable is replaced by log box-office revenue yields the same coefficient estimates except for that of log budget, which is increased by exactly 1. This is also the case for Column 2 and 3.

[12] See Wallace et al. (1993) and Prag and Casavant (1994). Notice that these studies use smaller samples and include the critical reviews as explanatory variables, which are unavailable before a movie is made.

Table 2: Spatial Regression of Log ROI

| | | (1) | (2) | (3) |
|---|---|---|---|---|
| Time | Constant | -0.807** | -0.572** | -0.505** |
| | Trend | 0.0021 | 0.0089** | 0.0107** |
| | Seasonality | 0.130** | 0.092** | 0.099** |
| Log Budget | | 0.098** | -0.0082 | 0.0091 |
| Genre | ... | Yes | Yes | |
| Rating | Restricted | -0.198** | -0.106** | |
| Crew | Actor | -0.0290 | 0.0640 | |
| | Director | -0.0644 | -0.0186 | |
| | Writer | 0.0817* | -0.0042 | |
| Spatial Lag | Log ROI | | 0.710** | 0.744** |
| $R^2$ | | 0.058 | 0.221 | 0.211 |
| $N$ | | 2,943 | 2,943 | 2,943 |

** Significant at the 95% level. * Significant at the 80% level. ROI is defined as the ratio between box-office and budget, both of which are normalized by CPI to be in 2014 million dollars. Dependent variable is the log ROI of the movies that started between 1995-2012 and has at least one precursor. A precursor here refers to any similar movie whose release date is earlier than that of the focal movie. Movies in 1975-1994 are used as possible precursors. Trend is the difference in years between the release date and the beginning of 1995. Seasonality uses a dummy for releases in Jun., Jul., Aug. and Dec. Genres are re-categorized into eight "big genres" to reduce the number of parameters. Actor is a dummy for movies with at least one leading actor that had previously taken a leading role in any of the top 5% grossing movies. Director and Writer are defined similarly. The spatial lag equals the average log ROI of the precursors.

Table 3: Polynomial Fit of Residual Size on Number of Precursors

| | Absolute Residuals |
|---|---|
| Intercept | 1.474 (.0615) |
| Number of Precursors | -0.174 (.0259) |
| Number of Precursors^2 | 0.0120 (.0032) |
| Number of Precursors^3 | -3.43e-4 (1.5e-4) |
| Number of Precursors^4 | 3.40e-6 (2.3e-6) |
| Average Partial Effect | -0.0751 |
| $N$ | 2,943 |

Numbers in the parentheses are standard errors. The dependent variable is the absolute value of the residuals from the last column of Table 2.

Table 4: Regression of the Number of Imitators / Precursors

| | | Log # of Precursors | Log # of Imitators |
|---|---|---|---|
| Time | Yearly Dummy | Yes | Yes |
| Genre | ... | Yes | Yes |
| Rating | Restricted | 0.0681* | 0.0979** |
| Crew | Actor | 0.142** | 0.0376* |
| | Director | 0.0837** | 0.0669** |
| | Writer | -0.0182 | -0.0571* |
| Log Budget | | 0.236** | 0.182** |
| Log ROI | | | 0.215** |
| $R^2$ | | 0.379 | 0.526 |
| $N$ | | 4,390 | 4,390 |

** Significant at the 95% level. * Significant at the 80% level. See Table 2 for some variable definitions. Here, an imitator of movie $j$ is defined as a movie that is similar to $j$ and started after $j$'s release. A precursor is a movie that is similar to $j$ and released before $j$'s start date. We add 1 to the number of imitators or precursors before taking the log. The entire sample (1975-2012) is included.

covariates.

The analysis above focuses on the conditional *expectation* of ROI. However, if one thinks of the ROI of a precursor as a noisy signal for the ROI of the focal movie, intuitively the *variance* of the prediction error should decrease with the number of precursors. To test this, in Table 3 we regress the absolute value of the residuals from the last column of Table 2 on the number of precursors. The estimated average partial effect is negative, indicating that the prediction error does become smaller with a larger number of precursors. The estimated polynomial also shows a diminishing decline rate, which is seen in standard Bayesian updating. Almost identical results can be obtained on the residuals from Column 2 in Table 2. An implication is that imitation can help reducing the uncertainty that firms have to face in bringing a potential movie to the box office.

Next we take a glimpse into the firm behaviors. Table 4 regresses the log number of "precursors" and "imitators" on various movie characteristics and log ROI. An imitator of a movie $j$ is defined as any similar movie that started production after the release of $j$. A precursor is any similar movie that was released before the start of $j$ (In this paper the definition of a precursor uses either the start dates or the release dates, depending on the context). Time dummies are added to control for the fact that the network is truncated outside the sample period. We see that movies with higher ROI are more likely to be imitated, supporting the conventional wisdom that there is firm-side learning in the movie industry.

A more subtle point in Table 4 is that bigger-budget movies tend to have more imitators as

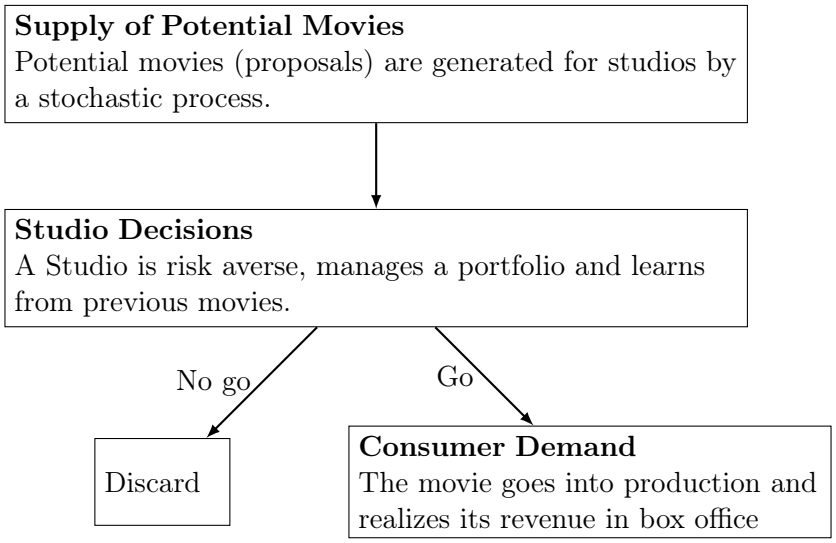



Figure 2: Model Overview

well as precursors. In other words, the network is denser amongst these movies. This suggests that studios rely more on imitation when it comes to the big-budgets, which is consistent with risk aversion. Budget size multiplies the risks that studios have to face in the box office for a potential movie. So for the big-budget movies it becomes particularly important to reduce the risks through imitation. On the other hand, small-budget movies are less risky and may actually favor novelty, a point which we will discuss more later in the paper.

# 4 Model

In this section we develop an empirical model of product entry for the industry. Figure 2 offers an overview of the model. First, there is the supply of potential movies where proposals arrive at the studios. We model this by a stochastic arrival process. Once the proposal lands on the desk of studio executives, they make a decision on whether to make the movie or discard it. After the movie is made, it goes to theaters where consumers decide how much box-office revenue it is going to receive.

There are two main differences from a standard entry model. First, we do not pre-fix a set of potential products but use an arrival process. From one perspective, this captures the finite supply of potential movies where not all conceivable movies are available at all times. From the perspective of modeling, it provides a pool of "baseline" products, against which we compare the set of produced movies to gauge the selectivity of firms. From a technical point of view, this also greatly reduces the dimension and complexity of the firm's investment problem, permitting a tractable model.

The second difference is that we allow for both risk aversion and learning on the firm side. It

means that a firm does not know for sure how well a potential product will be received by the market, but can learn from previous similar products to reduce that uncertainty. We use a network to keep track of the similarity amongst products. Risk aversion and learning also pose some challenges for the estimation, which we address in Section 5.

### 4.1 Consumer Demand

Throughout the model we denote by $x_j$ the vector of all the observed characteristics of movie $j$, which includes budget, genre, rating, crew power, production company, production start date and release date. We first model movie box-office performance. A movie typically stays in theaters for six to eight weeks, with the first 2 weeks collecting about 60% of the total domestic box-office revenues. Assume that consumer $i$'s utility from movie $j$ around the time of its release is

$$u_{ij} = U(x_j; \beta) + \mu_j + \varepsilon_{ij}.$$

Of course, $U(\cdot)$ need not incorporate all the elements of $x_j$. For example, it is probably far-fetched to argue that the production company or production start date would enter consumer utility.[13]

$\mu_j$ is an unobserved component which we refer to as the "latent quality." It captures the average consumer tastes over the unobserved characteristics of the movie at the time of release. These characteristics include visual effects, theme, storyline, narrative method, pace, music, and so on. To the extent that similar movies share these characteristics, the $\mu$'s of similar movies should be correlated. The correlation is an essential part of the model that we will address when describing the supply side.

Suppose that the individual consumer chooses between going to a movie theater to watch $j$ and an "outside option" with utility:

$$u_{i0} = \varepsilon_{i0}.$$

Then, assuming type-I extreme value distribution for the idiosyncratic errors $\varepsilon_{ij}$ and $\varepsilon_{i0}$, we have the "market share" of $j$ given by $1/(1 + e^{-U(x_j,\beta)-\mu_j})$. To convert market share into revenue, we multiple it by the market size and average ticket price at theaters.[14] With the multiplier at time $t$ denoted by $m_t$ and the release date of $j$ denoted by $r_j$, the box-office revenue for $j$ can be expressed as

$$\pi_j = m_{r_j}/(1 + e^{-U(x_j,\beta)-\mu_j}), \tag{1}$$

[13]The model abstracts away from several factors that affect demand, including the marketing expenditure, timing of the release and number of screens. These factors are determined after the movie is made, and are endogenous outcomes of budget size, movie quality and competition at theaters. See Hennig-Thurau et al. (2006) for the relative importance of marketing vs. movie quality. They find that overall, quality is more important. See Elberse and Eliashberg (2003), Ainslie et al. (2005) and Einav (2007, 2010) for the exhibition dynamics, competition, and release timing.

[14]We treat $m_t$ as an exogenous time series. It is an known fact (as well as a puzzle) that theatrical ticket price hardly varies across seasons and movies. See Orbach and Einav (2007) for more discussions.

The market size is taken as the population of "moviegoers" who go to cinema at least once a year, about two-third of the entire population.[15] Note that there is a one-to-one relation between the box-office revenue and latent quality of a movie. So even if $\mu_j$ is unknown before the release of $j$, it is revealed after it is exhibited in the theaters.

### 4.2 Arrival Process

Treating time as continuous, we let potential movies arrive at a Poisson rate $\eta_f$ for firm $f$.[16] Suppose that a potential movie $j$ arrives at time $t$ for firm $f$. If it ever gets produced, we record its arrival time by $a_j$ and its production period is $(a_j, r_j]$. The observed characteristics $x_j$, latent quality $\mu_j$, as well as the similarity between $j$ and the existing movies, are drawn from a state-dependent distribution. The state is denoted as $\mathcal{J}_t$ and is the collection of the observed characteristics, latent qualities and similarity network of all the existing movies at $t$ (released or in-production).

For a "shell proposal" that has just arrived, we can either (i) determine its characteristics first and then determine its similarity with the existing movies, or (ii) determine the similarity first and then determine its characteristics conditional on the similarity. For the model to be empirically tractable, the second procedure is more appropriate because many characteristics are not observed in the data after all. In addition, it closely captures the associative nature of the creation process where new ideas are based on old ones (see, for example, Mednick (1962), Weitzman (1998) and Uzzi (2013)). One can think of the scenario where some of the crew or technology in a previous movie is deployed in a new but similar movie.

In principle, one can specify distribution for all the observed characteristics. However, our reduced-form analysis has indicated that the similarity measure incorporates much of the proximity with respect to genre, rating and crew power. Later in Section 6, we also show that these covariates add little to the prediction of market shares in the demand model. For parsimony, on the supply side we let $x_j$ include only the budget and covariates related to the release time (seasonality and a time trend). It is conceptually simple to extend the model to include all the covariates.

Before continuing to the specification of the arrival distribution, we want to mention that the arrival process is a latent structure. We have experimented with many variations of the process and our choice has been guided by both economic intuition and patterns in the data. The extent to which the model is capable of reproducing the data is partially assessed in Section 6.2.

[15] See *Theatrical Market Statistics*, MPAA.

[16] The model allows a different arrival rate for each firm, which introduces quite a lot parameters. For estimation, I use a single arrival rate and assign each arrival to a firm with the probability proportional to the number of movies that belong to it in the data. For each movie, the first listed production company is counted as the firm for that movie.


Solid nodes represent existing movies. The arrival, represented by a hollow node, attaches to each existing movie independently in the first stage. A realization is displayed on the left where it attaches to node (a). Given this outcome, the arrival further attaches to each neighbor of (a) in the second stage. A realization is displayed on the right where it attaches to (b).

Figure 3: An Illustration of the Two-Stage Process

**Similarity** The set of existing movies at time $t$, $\{k : a_k < t\}$, includes those that are either released or still in production. The similarity amongst them is described by a network. Given our data, we use a simple network where a link represents that the two movies are similar. Ideally one would like to use a weighted network where a link between two movies is assigned a weight that reflects the degree of similarity between them. Weighted similarity networks is a topic left for future research.

The similarity between $j$ and the existing movies is formally described by which existing movies become linked to $j$. In the language of the evolving network models, the arrival "attaches" to the similarity network at time $a_j$. We use $y_j$ to denote the outcome of attachment process. It is a vector of the length of the number of existing movies, where $y_{k,j} = 1$ indicates that a link is formed between $j$ and the existing movie $k$, and $y_{k,j} = 0$ otherwise. We let the attachment probability follow a logit model:

$$\Pr(y_{k,j} = 1|\mathcal{J}_t) = 1/(1 + e^{-F(x_k,t;\gamma)}), \quad \forall k. \tag{2}$$

We specify $F(x_k, t; \gamma) = \gamma_0 + \gamma_1[k \in f] - \gamma_2(t - a_k)$. The first term is a constant. The second term is an indicator dummy that gives potentially higher probability to attachment towards the movies produced by the same firm $f$. It captures the possibility that a firm favors its own types of movies. The third term discounts movies by their arrival dates, capturing the idea that movies gradually become obsolete and unlikely to be imitated anymore. These two specifications are also consistent with the properties of the observed network that we reported in Table 1.

Another important property of the observed network is substantial clustering. A set of nodes is said to be clustered if they are densely linked with each other. In social networks, clustering refers to the property that "my friends are friends themselves." The average clustering coefficient[17] for

[17]The clustering coefficient of a node equals the number of triangles that it belongs to divided by the number of triangles that it would belong to if all of its neighbors were linked with each other. See Watts and Strogatz (1998).

our similarity network is 0.22. As a comparison, randomly assigning the same number of links to the same number of nodes yields a clustering coefficient typically less than 0.01. Underlying the clustering is the transitivity of similarity: if $j$ is similar to $k$ and $k$ is similar to $\ell$ , then it is more likely that $j$ and $\ell$ are similar as well.

We model the clustering through a two-stage attachment process. Similar ideas has been used in Holme and Kim (2002) and Jackson and Rogers (2005) to develop simple but flexible ways to generate clustering in social networks. Specifically, in the first stage, $j$ forms link with each existing movie independently. In the second stage, for each $k$ linked in the first stage, $j$ further forms links with each neighbor of $k$ with probability $\omega$. An example of the two-stage process is illustrated in Figure 3. In the Appendix we show how to calculate the 1st-stage probabilities from (2).

**Observed characteristics** We specify a distribution $\Pr(x_j|y_j, \mathcal{J}_t)$ from which the observed characteristics of $j$ can be drawn. Aside from the state $\mathcal{J}_t$, the distribution also depends on $y_j$. This allows the observed characteristics of the arrival to be correlated with those of its precursors, $\{k : y_{k,j} = 1\}$. For example, an arrival that is similar to a group of big-budget movies should be more likely to have a big budget as well. Had we included genre on the supply side, an arrival whose precursors are mostly science fictions should be more likely to tell futuristic stories such as hacking dreams and wormhole travel.

Specifically, the budget for $j$ will be drawn from a truncated normal distribution. The coefficient of variation, which we denote by $\chi$, is to be estimated as a parameter. If the set of the precursors for $j$ is nonempty, the mean of the truncated normal is set to the average budget of the precursors. Otherwise the mean equals $\theta$, which is another parameter to be estimated.[18] The release time is determined by computing the production period $|r_j - a_j|$ as a nonparametric function of the budget, which is estimated "off-line" with the data on production start date. This is a simplification and we abstract away from the factors that influence the release date after the green-light decision.

**Latent quality** We want to consider two factors in specifying $\Pr(\mu_j|x_j, y_j, \mathcal{J}_t)$. First, as with the observed characteristics, we want to allow correlation between $\mu_j$ and $\mu_k$ whenever $k$ is a precursor of $j$. Second, recall that latent quality refers to the consumer tastes *at the time of* the movie's release, so to the extent that consumer tastes are time-varying, we want to allow the difference in the release time, $|r_j - r_k|$, to dilute the correlation between $\mu_j$ and $\mu_k$. Thus, a

[18] In the data, budget distributes around the mean budget of precursors in a truncated normal shape, and the dispersion hardly shrinks with the number of precursors. We use $[1, 350]$ as the truncation interval, as the biggest budget observed in the data is \$343m. Results are not sensitive to the choice of the interval upper bound.

suitable candidate is

$$\Pr\left(\mu_j | x_j, y_j, \mathcal{J}_t\right) = \mathcal{N}\left(\frac{\lambda \sum_{k\sim j, a_k<t} \phi^{|r_k - r_j|} \mu_k}{1 + \lambda \sum_{k\sim j, a_k<t} \phi^{|r_k - r_j|}},\ \frac{\sigma^2}{1 + \lambda \sum_{k\sim j, a_k<t} \phi^{|r_k - r_j|}}\right), \tag{3}$$

where $\phi < 1$ is a discounting factor, and $k \sim \ell$ denotes that there is a link between $k$ and $\ell$. Note how (3) has the intuitive form of standard Bayesian updating under normality, where the latent quality of a precursor can be thought of as a signal for the latent quality of the potential movie, and the information of the signal is discounted by the age of the precursor. Parameter $\lambda$ can be thought of as a measure of similarity. For $\lambda = 0$, the latent qualities are independent as if there is no similarity between movies. Parameter $\phi$ measures the inertia of consumer tastes. For the extreme case $\phi = 0$, consumer tastes change so rapidly that two movies released at different times, even if similar, will have completely uncorrelated market receptions. As we will show in model fit (Section 6.2), normality seems a reasonable assumption on the latent quality.

### 4.3 Product Portfolio

When it is the time to green-light a potential movie, studio sees the story and screenplay, and in most cases has a reliable estimate of the budget and release date. The producer often has secured some of the crew and is aware of who else she or he needs to recruit. However, much more uncertainty remains on how this particular movie will be received at the box office.[19] Our corresponding modeling assumption is that firms do not observe the latent qualities of the yet to be released movies (either in-production or just arriving), which they need to form beliefs about.

To formally model the decision making process, we start with the information set for the firms, denoted as $\mathcal{I}_t$. The set includes the observed characteristics and the network of the existing movies as well as the arrival, if there is one at $t$. It also includes the latent qualities of the *released* movies. Given the one-to-one relation between the box-office revenue and latent quality in (1), it is equivalent to include $\pi_k$ instead of $\mu_k$ in $\mathcal{I}_t$ for each released $k$. Notice the important difference between $\mathcal{J}_t$ and $\mathcal{I}_t$ that the later does not contain the latent qualities of the in-production movies, which are still unobserved at time $t$.

Given this information set, we can work out the belief of the firms. For a single arrival $j$ that is only linked with released movies, the posterior belief on its latent quality $\Pr(\mu_j|\mathcal{I}_t)$ is simply given by (3). However, for the general case $\Pr\left(\{\mu_k : a_k \leq t, r_k \geq t\}|\mathcal{I}_t\right)$, one needs to be more careful. For example, it is possible for a movie to arrive after but be released before another movie. In such case, one can learn about the second movie from the first one. One can also learn from indirectly linked movies when the direct neighbor has not been released. In addition, the

[19] Here is a description of the "green-light" process by a senior studio executive: "We bring together all studio department heads. [The production costs] is our most reliable estimate, and that thus forms the basis for our launch decision.... In the end; ... Someone in the meeting has to put his or her reputation on the line and say 'yes' - regardless of whether the numbers add up" Eliashberg et al. (2006)

network structure implies correlations among the latent qualities. In the Appendix we derive a closed-form expression for the general posterior belief.

Given the belief of the firms, we can now model their investment decisions. Given $\mathcal{I}_t$, we denote by $P_{f,t} \equiv \{k \in f : a_k < t, r_k \geq t\}$ the set of $f$'s in-production movies, which we can think of as $f$'s portfolio. Suppose that there is an arrival movie $j$ for the firm at time $t$. We will ask the firm to decide if it is desirable to add $j$ into its portfolio. The present value of $P_{f,t}$ is

$$\Pi(P_{f,t}) = \sum\nolimits_{k \in P_{f,t}} \delta^{r_k - t} \pi_k,$$

where $\delta$ is a discounting factor and $\pi_k$ is the box-office revenue of movie $k$ given by equation (1), and in particular, depends on $\mu_k$. If $j$ is accepted, the present value of the new portfolio becomes

$$\Pi(P_{f,t} \cup \{j\}) = \delta^{r_j - t} \pi_j + \Pi(P_{f,t}),$$

where $\pi_j$ depends on the latent quality $\mu_j$. At time $t$, the firm is uncertain about these values because it does not know the $\mu_j$ and $\mu_k$'s for $k \in P_{f,t}$. We allow the firm to be risk-averse so it evaluates the present values with a concave function $V$. We specify $V$ to take the utility form of constant absolute risk aversion (CARA). The firm accepts $j$ iff

$$\mathbb{E}\left(V\left(\Pi(P_{f,t} \cup \{j\}) - b_j - \zeta_j\right) \middle| \mathcal{I}_t\right) > \mathbb{E}\left(V(\Pi(P_{f,t})) \middle| \mathcal{I}_t\right), \tag{4}$$

where $b_j$ is the production budget of $j$ and $\zeta_j$ is an independent decision error that captures the factors unobserved to us but known to the firm. We specify $\zeta_j = (e^{z_j} - 1)b_j$ where $z_j$ is distributed $\mathcal{N}(0, \rho^2)$. The firm discards the arrival if condition (4) does not hold.

We can readily define a risk-free equivalence of the revenue $\pi_j$, denoted by $\overline{\pi}_j$, through the equation

$$\mathbb{E}\left(V(\Pi(P_{f,t} \cup \{j\}))|\mathcal{I}_t\right) = \mathbb{E}\left(V(\Pi(P_{f,t}) + \overline{\pi}_j)|\mathcal{I}_t\right).$$

When $P_{f,t}$ is empty, it reduces to the more familiar $\mathbb{E}\left(V(\pi_j)|\mathcal{I}_t\right) = V(\overline{\pi}_j)$. The definition allows us to state condition (4) alternatively as $\overline{\pi}_j - b_j - \zeta_j > 0$. We may also view $\overline{\pi}_j / b_j$ as the risk-adjusted ROI of movie $j$, and a movie is accepted iff its log risk-adjusted ROI is larger than $z_j$. So from the perspective of the econometrician, the acceptance probability follows a probit model with respect to the log risk-adjusted ROI.

The journey of a movie often goes beyond production and domestic box office, spending more on advertising and exhibition, while earning more from home video sales and international markets. However, these additional revenues and expenses are heavily influenced by the budget and domestic box office. To capture them, we may specify that the risk-adjusted ROI is $c\overline{\pi}_j / b_j$ where $c$ is a coefficient. A larger coefficient increases the acceptance probability for all the arrivals. Notice that, with a latent arrival process, the model is observationally equivalent if we halves the acceptance probability for every type of arrivals and double the arrival rate. In the Appendix, we

use a simplified version of the model to explain why the coefficient can only be weakly identified.

An important feature of our formulation of firm's decisions is that it takes into account the correlations between the box-office revenues of different movies. To the extent that the firm is risk averse, it would like to "diversify" its portfolio and avoid investments in many similar movies at once. However, the formulation treats the firm myopic, not taking into account how a decision today will affect future arrivals and decisions. To solve for a full model of forward-looking decisions with the network similarity structure is beyond this paper, but is a challenging topic for future research.

# 5 Model Estimation

## 5.1 Demand

Though firms select what products to introduce to the market, most applications estimate demand by assuming that the set of products is exogenous and focus on other sources of endogeneity (e.g., price). Even in studies of market or product entry, it is standard to retain exogeneity in the unobserved component (in our context, the $\mu$) by arguing that firms have no knowledge of it before entry (see, for example, Aguirregabiria and Ho (2011) and Eizenberg (2014)). Because our model relaxes this assumption by allowing firms to learn about the $\mu$ before entry, it requires an extension of the standard estimation technique.[20]

To be more specific, first note that the following regression equation can be directly obtained from the box-office equation (1):

$$\log(\pi_j) - \log\left(m_{r_j} - \pi_j\right) = U(x_j, \beta) + \mu_j. \tag{5}$$

Due to endogenous entry, here the standard moment condition $\mathbb{E}(\mu_j|x_j) = 0$ generally does not hold. For example, $\mu_j$ can be positively correlated with the budget $b_j$ in $x_j$, because a bigger budget implies larger risks which typically need to be compensated by a higher belief on $\mu_j$ for entry.

We solve this problem by controlling for what firms can learn about $\mu_j$ at the time of entry. In our model, $(x_j, y_j, \mathcal{J}_{a_j})$ contains the firm's information set at time $a_j$. It determines the arrival distribution of of $\mu_j$ through (3). Let $\xi_j \equiv \mu_j - \mathbb{E}(\mu_j|x_j, y_j, \mathcal{J}_{a_j}; \beta, \lambda, \phi)$, which is the difference between the realized latent quality and the mean of its arrival distribution. Then we have $\mathbb{E}(\xi_j|x_j) = 0$. This actually constitutes the first set of our moments. Identification of parameter $\sigma$ requires us to match the dispersion of the arrival distribution as well, so we define a second difference: $\iota_j \equiv \xi_j^2 - \mathbb{E}(\xi_j^2|x_j, y_j, \mathcal{J}_{a_j}; \beta, \lambda, \phi, \sigma)$. Our demand-side estimation are then

[20] An important difference here from the standard spatial econometrics is that the network is not exogenous. For general treatment of spatial econometrics, see Bradlow (2005) and LeSage (2008).

based on the following mean-independence moments:

$$\mathbb{E}\left[(\xi_j, \iota_j)\,|x_j, y_j, \mathcal{J}_{a_j}\right] = 0.$$

Moments like these leave us with many instruments to choose from the conditioning set to interact with $\xi_j$ and $\iota_j$.[21] To identify $\beta$, we interact $\xi_j$ with $x_j$. To identify parameter $\lambda$ and $\phi$, we interact $\xi_j$ with the average latent quality of the precursors for $j$, and the average latent quality of the precursors that were released several years earlier than $r_j$. To identify $\sigma$, we interact $\iota_j$ with a constant term and the number of the precursors for $j$. To the extend that the $\mu$'s are unobserved, we compute them through equation (5) as a function of the data and unknown parameter $\beta$.

The sample moments average across the movies that started production after 1995. The movies in 1975-1994 are counted as possible precursors of these movies. Not conditioning on this initial sample should not affect the asymptotics of the estimator as the sample period expands, but is likely to cause sizable finite-sample bias. Finally, because the estimation requires a numerical search jointly over $(\beta, \lambda, \phi, \sigma)$, it can be computationally intensive when we include many covariates in $x_j$. However, we can save computational time by using the OLS estimates of (5) as the initial parameter guess for $\beta$.

### 5.2 Supply

The estimation procedure for the supply side is relatively straightforward. Essentially, we match the properties that the model predicts for the *produced* movies with those observed in the data. Specifically, index the movies in data by arrival date so that $j$ is the first movie that arrives after $j-1$. The full history up to time $a_j$ can be summarized as $(x_j, y_j, \mu_j, \mathcal{J}_{a_j})$. Let $H$ be a function this history. For notation, we write $H(x_j, y_j, \mu_j, \mathcal{J}_{a_j})$ as $H_j$. The specification of $H$ depends by the moments that we want to match on the supply side to identify the parameters. We give the specification of $H$ below after discussing the identification.

Collect the supply-side parameters in $\Lambda$. For any value of $\Lambda$, given the state at time $a_{j-1}$, our model makes a prediction of $H_j$ with the error of prediction given by

$$h(x_j, y_j, \mu_j, \mathcal{J}_{a_j}; \Lambda) \equiv H_j - \mathbb{E}(H_j | x_{j-1}, y_{j-1}, \mu_{j-1}, \mathcal{J}_{a_{j-1}}; \Lambda).$$

The conditional expectation does not have closed forms, but can be evaluated through simulations. This evaluation step is computationally intensive and required us to make use of a

[21] For reasons why we do not use many moment conditions, see Andersen and Sørensen (1996) and more recently Han and Phillips (2006).

computer cluster.[22] Our supply-side estimation relies on the following moment conditions:

$$\mathbb{E}\left(h(x_j, y_j, \mu_j, \mathcal{J}_{a_j}; \Lambda)\right) = 0.$$

The estimate of $\Lambda$ is obtained following the procedure of the Generalized Method of Moments. It searches for the parameter values that minimizes a norm of the sample counterpart of the moment conditions: $||\frac{1}{n-k+1}\sum_{j=k}^{n} h(\cdot)||$ where $k$ is the first movie produced since 1995. Again, movies in 1975-1994 are counted as possible precursors but not included in sample moments.

In our model, the set of produced movies is a joint outcome of both the arrival process and the production decisions. Here we provide some intuition as to how these two parts can be separately identified. In the Appendix, identification is shown for a simplified version of the model. For the full model, we show through Monte Carlo experiments that the parameter values can be recovered with reasonable precision.

Suppose that the decision parameters were known to us. Then the identification of the arrival parameters would be fairly straightforward. The arrival rate can be identified by the frequency at which movies are produced. The attachment parameters $\gamma$ can be generally identified by the properties of the precursors. The coefficient of variation for the arrival distribution of budget, $\chi$, can be identified by the variation of budget in the data. The mean budget for an arrival without precursors, $\theta$, can be identified by the average budget size of the movies without precursors; it can also be identified simply by the proportion of the movies without precursors in the data, because $\theta$ affects the probability with which arrivals without precursors are accepted.

Additional moments are required to identify the decision parameters. In the reduced-form analysis, we have observed that big-budget movies are more imitative and viewed it as a suggestive evidence of risk aversion. In line with this interpretation, we identify the coefficient of risk aversion by the difference in the degree of imitation between big-budget and small-budget movies. Given the model assumption that the mean budget of an arrival equals the mean budget of the precursors, risk aversion can also be identified simply by the average budget size of the produced movies. This is because the acceptance probability for big-budget movies decreases with the level of risk aversion. The other decision parameter is the size of the decision error, $\rho$. It can be identified by the average latent quality or ROI of the produced movies, as a smaller error size implies that the firms are more selective in accepting arrivals. Another way to identify $\rho$ is simply to look at the proportion of movies produced with risk-adjuste ROI less than 1.

The specification of $H$ follows our identification argument. We include in $H_j$ the time elapsed since last movie production: $a_j - a_{j-1}$, an indicator whether there are precursors for $j$, the log number of the precursors, and the log number of the triangles created in the attachment of $j$. We also include, within the precursors of $j$, the proportion of the movies produced by the same firm

[22] It is not necessary to use a very large number of simulations to evaluate one conditional expectation, as the simulation errors are averaged across the movies. We chose to use 100 simulations. Nevertheless, one evaluation of the objective function takes around 10 mins on a quad-core desktop.

Table 5: Model Parameter Estimates, Demand-Side

| Parameters | | I | II | III |
|---|---|---|---|---|
| Time | Constant | -7.14 (.14) | -7.24 (.18) | -7.27 (.17) |
| | Trend | -0.0202 (.004) | -0.0200 (.008) | -0.0166 (.008) |
| | Seasonality | 0.151 (.05) | 0.110 (.04) | 0.121 (.04) |
| Budget | | See Fig. 4 | See Fig. 4 | Yes |
| Rating | Restricted | -0.198 (.05) | -0.209 (.06) | |
| Genre | Drama | 0 | 0 | |
| | Comedy | 0.418 (.07) | 0.230 (.08) | |
| | Action/War | 0.197 (.07) | 0.187 (.08) | |
| | Family | 0.385 (.09) | 0.433 (.15) | |
| | Sci-Fi/Advent. | 0.358 (.1) | 0.308 (.09) | |
| | Horror | 1.11 (.1) | 0.970 (.2) | |
| | History/Bio. | 0.107 (.1) | 0.122 (.1) | |
| | Others | -0.426 (.4) | -0.686 (.3) | |
| Crew Power | Actor | -0.0264 (.05) | 0.0139 (.05) | |
| | Director | -0.0677 (.06) | -0.0282 (.06) | |
| | Writer | 0.121 (.06) | 0.0531 (.06) | |
| Similarity ($\lambda$) | | | 0.529 (.07) | 0.583 (.07) |
| Disc. Factor ($\phi$) | | | 0.929 (.02) | 0.925 (.02) |
| Std. Dev. ($\sigma$) | | | 1.82 (.05) | 1.87 (.05) |
| $R^2$(share) | | 0.557 | 0.658 | 0.653 |
| $R^2$(ROI) | | 0.0716 | 0.283 | 0.273 |

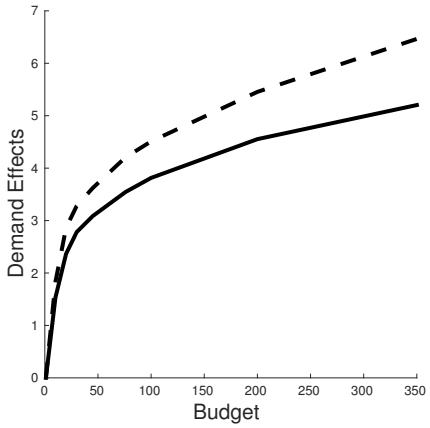
Column I displays the OLS estimates of equation (5). Column II and III display the GMM estimates. See Table 2 for definitions of some of the variables. The utility of budget is estimated as a piecewise linear function; see Figure 4 for the estimates. Discounting factor $\phi$ is yearly. The numbers in parentheses are standard errors. $R^2$(share) measures the prediction for the log market share. $R^2$(ROI) measures the prediction for the log ROIs.

as $j$ and the proportion of the movies started several years earlier than $a_j$. Finally, we include the log budget of $j$, the log budget squared, the latent quality $\mu_j$, and an indicator whether $\overline{\pi}_j < b_j$.

# 6 Estimation Results

## 6.1 Parameter Values

**Demand side** Table 5 displays the estimates of the demand-side parameters. Specifically, column I displays the estimates from an OLS regression of the revenue equation (5); column II displays the GMM estimates with all the covariates; column III displays the GMM estimates




The solid curve are the GMM estimates of the piecewise function for the utility from budget, corresponding to Column II in Table 5. The dashed curve are the OLS estimates corresponding to Column I. The shape displays diminishing marginal utility. However, the diminishing rate is slower than that of a logarithm specification.

Figure 4: Estimated Effects ($\beta$) of Budget

with only the covariates related to budget and release time.

First notice that the estimates do not differ too much across the three configurations, so some common observations can be made. There is a small but statistically significant downward trend, which may be attributed to the growth of the home video market as an alternative to movie theaters. The demand for movies tends to be higher in the summer and at the end of the year, which is consistent with the results in Einav (2007). As expected, a "restricted" MPAA rating reduces demand. Interestingly, Horror movies are the best bet for studios to make profits.[23] The effect of star power is insignificant, which is consistent with our reduced-form analysis and the finding in Ravid (1999) that stars capture their economic rent. Finally, the effects of budget are estimated as a piecewise linear function and plotted in Figure 4. The shapes of the function exhibit diminishing marginal utility.

Under the estimates of $\lambda$ and $\phi$, our model implies that for the movies in the data, learning on average reduces the variance in $\mu_j$ at the time of arrival $a_j$ by more than 60%. The estimate of $\phi$ indicates a quite rapid change of consumer tastes over time. For example, in updating on the latent quality of a potential movie, a 10-year old precursor counts less than half as much as a precursor from the last year. The estimate of $\sigma$ implies an enormous uncertainty in the latent quality. To see the magnitude, recall that $\sigma$ is the standard deviation of the latent quality for an original movie (without any precursors); one standard deviation equals about the effect of raising the budget of a \$10m movie to \$65m, or the budget of a \$100m movie to over \$300m.

The difference between the GMM and OLS estimates of the effects of budget (Figure 4) can be explained by endogenous entry. For example, for a big-budget movie to be produced, a high

[23]For a stimulating discussion on this, see "Let's Get Scared: Why Horror Movies Are Immune to the Digital Onslaught." September 16, 2013, *Yahoo Movies*.

Table 6: Supply-Side Parameter Estimates

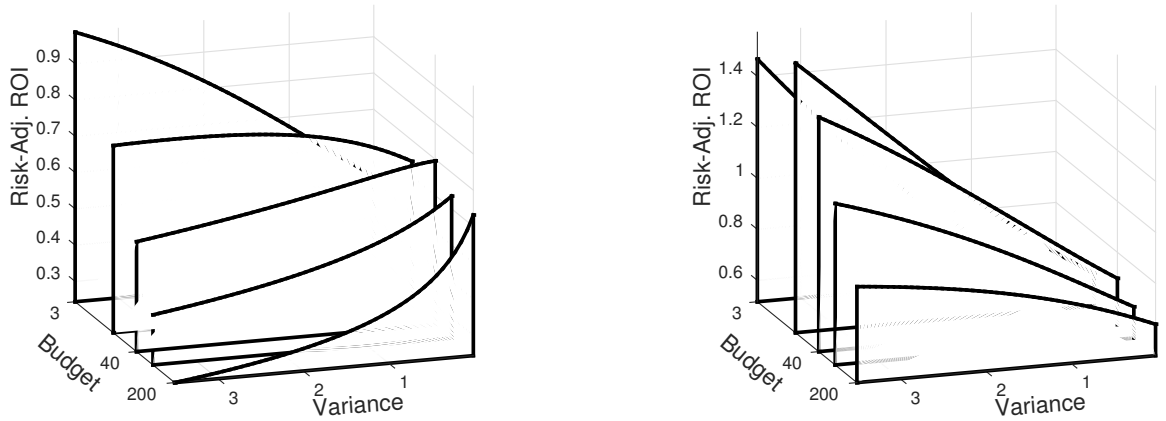
| Parameters | | Estimates | |
|---|---|---|---|
| Attachment | 2nd-stage Probability ($\omega$) | 0.219 | (.003) |
| | Intercept ($\gamma_0$) | -4.711 | (.03) |
| | Own Movies ($\gamma_1$) | 1.695 | (.06) |
| | Time Difference ($\gamma_2$) | -0.230 | (.006) |
| Obs. Characteristics | Budget Mean without Precursors ($\theta$) | 56.9 | (10) |
| | Budget Coeff. of Variation ($\chi$) | 1.586 | (.08) |
| Coeff. of Risk Aversion ($\alpha$) | | 0.0201 | (.0047) |
| Std. Dev. of the Normal Shock ($\rho$) | | 0.570 | (.05) |
| Yearly Arrival Rate ($\eta$) | | 622 | (31) |

For $\gamma_2$, time difference is expressed in years. Budgets are expressed in 2014 million dollars. A single arrival rate is estimated where any arrival is assigned to one firm according to the empirical distribution of movie ownership. The firm discounting factor $\delta$ is set at .975. Numbers in the parentheses are standard errors computed by parametric bootstrapping (see Appendix).

belief on $\mu$ is typically required to compensate the associated large risks. This introduces a positive correlation between $b_j$ and $\mu_j$, making the OLS estimates biased towards larger effects of budget. The estimated effects of genre, rating and crew power tend to be smaller with the GMM (Table 5). This is because these effects are incorporated to certain extent by the latent qualities of the similar movies.

As to the explanatory power, by accounting for the similarity network, Column II increases the $R^2$ from .56 to about .66 when compared to Column I. Notice that the model is intended to explain the market shares, so budget size is a major predictor and contributes significantly to the $R^2$. In terms of explaining the ROI, the model in Column I performs very poorly, while the model in Column II provides a substantial improvement. This is in line with our reduced-form results (Table 2).

As we move from Column II to III, the $R^2$ remains almost identical, and $\lambda$ picks up some of the effects of the dropped covariates. This is again in line with our reduced-form analysis, suggesting that (i) the performances of the precursors are much better predictors than genre, rating and crew power, (ii) the similarity measure has incorporated the proximity in these covariates to certain extent. The result justifies us to not consider these covariates on the supply side, making the model and estimation much more tractable.




The plots show the risk-adjusted ROI, $\overline{\pi}_j/b_j$, of a hypothetical movie $j$ as a function of the budget size and variance in the latent quality $\mu_j$. The latent qualities of the precursors for $j$ are all assumed to be $\sigma/3$. Budget is again expressed in 2014 million dollars. The plot on the left uses model estimates, while the plot on the right takes $\alpha \to 0$ so firms are risk-neutral. ROI only takes domestic box office and production budget into account.

Figure 5: Risk-adjusted ROI against Variance in $\mu$ for Various Budget Sizes

**Supply side** Table 6 displays the supply-side estimates. First shown are the parameters pertaining to the attachment process. We see a sizable 2nd-stage attachment probability, which is consistent with the degree of clustering in the observed network. Estimates also indicate that there is a much higher probability for an arrival for a firm to attach to the movies produced by the same firm. This reflects the fact that a studio may develop "tastes" for certain types of movies over time, and further, may have signed exclusive contracts with the crew from its past releases.

The coefficient of absolute risk aversion, $\alpha$, is estimated to be both statistically and economically significant. In Figure 5, the graph on the left plots risk-adjusted ROI as a function of the budget size and variance in the $\mu$ under the estimated $\alpha$; the graph on the right plots the risk-neutral case where $\alpha \to 0$. The two graphs depict every different preferences over novel vs. imitative movies, indicating that risk aversion does play a significant role.

Further, from the left graph in Figure 5, we see that risk-adjusted ROI decreases with the variance in $\mu$ for big-budget movies. This is expected because a higher variance in $\mu$ implies a higher level of risks. However, what comes at a surprise is that this relation is reversed for small-budget movies. This is because the box-office revenue distribution for a small-budget is very right-skewed. A larger variance in $\mu$ expands the right tail but not as much for the left tail, which is bounded by zero, resulting in an increase in the expectation of the box-office revenue. In other words, for small-budgets, novelty increases the chance of becoming a big hit. This mapping of risk-adjusted ROI is a driving force underlying some of the predictions later in the counter-factual analysis.[24]

[24]Given this mapping, one may ask the question why not split the money for a big-budget into many small-

Recall that we allow for a decision error on the firm side, capturing factors not observed by us but known to the firms. The larger is the size of the error, $\rho$, the less predictive is our model on the production decisions. In the extreme case where $\rho \to +\infty$, all types of arrivals are accepted with equal probability. Under the estimate of $\rho$, the acceptance probability ranges from about .01 to .51 for the range of the risk-adjusted ROIs in the left graph of Figure 5. So our model captures a good deal of the production decisions. Finally, the estimate of the arrival rate implies that around three quarters of the arrivals are rejected. Because the arrival process is not designed to capture the supply of movie scripts, the estimate should not be interpreted literally as the acceptance rate of movie scripts.[25]

### 6.2 Model Fit

To make an assessment on model fit, we simulate the model from 1995 all the way to 2012 conditional on the initial data from 1975 to 1994. In Figure 6, we compare the simulated data to the real data. Given the important trade-off between budget size and uncertainty in the firm's decision (see Figure 5), we look at the scatter plots of the number of precursors against the budget, the distributions of the number of precursors, and the budget distributions. In the last row of the Figure we also show the distributions of ROI.

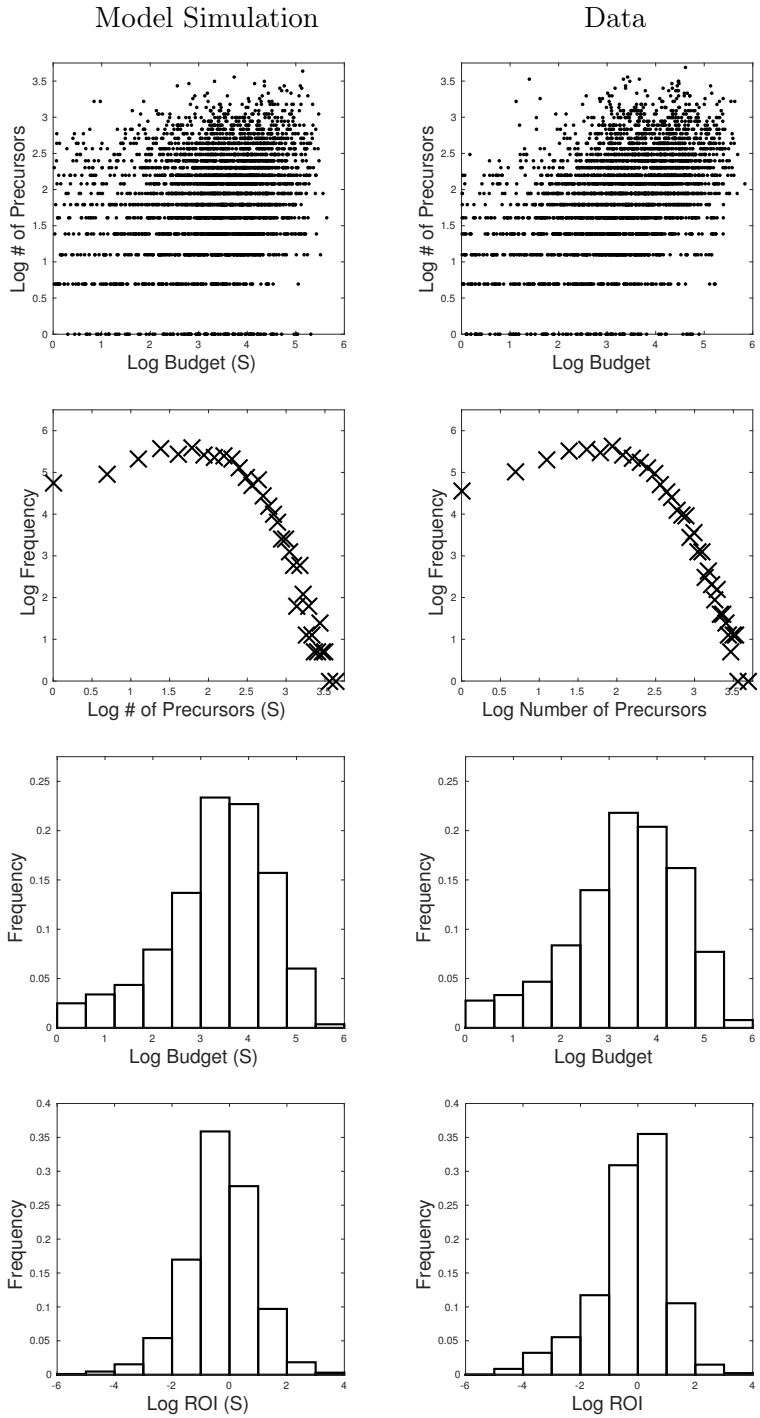
Considering that there are fewer than ten parameters on the supply side, the model does a satisfying job reproducing the patterns in the data. Because the production and release strategies can be different across movies with diverse sizes, production companies and release years, it is difficult for the model to capture all patterns in the data. For example, the model seems to under-produce the very big-budget movies. This could be caused by risk aversion heterogeneity across firms, which our model fails to capture. Blockbusters are often produced by major studios that are financially more capable than independent production companies. The model also seems to produce a smaller left tail for the log ROI. This could be caused by the normality assumption on $\mu$. The fatter left tail in the data suggests that it may be better to use a distribution that allows some degree of negative skewness. Enriching the model for a better fit with the data is left for future research.

## 7 Counter-factual Experiments

In this section we use several counter-factual experiments to provide further insights on how learning affects product entry. First, we show that learning matters by examining what happens

budget and novel movies? The immediate answer is that movie supply is not infinite. Once there are many small-budget movies produced, it becomes difficult for another original small-budget to arrive. Goettler and Leslie (2005) asked the same question and offered a few alternative explanations.

[25] For readers interested in the transaction of movie scripts, see Luo (2014). However, rejected scripts are not included in her data and a rejection rate is not provided.




The model is simulated for once from 1995-2012 conditional on the data up until 1995. The column on the left plots the simulated data, while the column on the right plots the real data. Each row plots, respectively: (i) the log number of precursors against the log budget, (ii) the frequency of the log number of precursors, (iii) the histogram of log budget distribution, (iv) the histogram of log ROI. A precursor for $j$ is any $k$ that satisfies $k \sim j$ and $a_k < a_j$. We add 1 to the number of precursors before taking its log.

Figure 6: Comparisons Between Data and Simulation

if firms stop learning. In addition, we try to quantify the indirect effects of learning as firms imitate each other's products. Second, we examine the consequences of an increase in the arrival rate, and provide an explanation to the increasing budget size in the industry. Third, we examine what happens when there is a change in firm risk attitude, and provide an explanation to the rise of imitation in motion pictures.

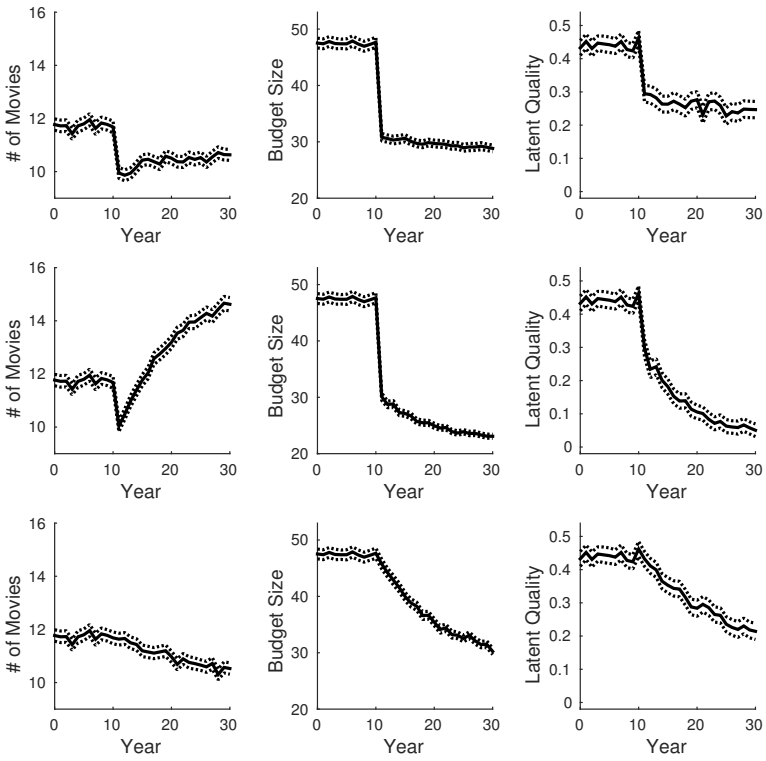
For all the counter-factuals, we introduce the changes at the steady state of the model. For the model to have a steady state, we remove the demand trend and set both the market size and ticket price constant over time. The rest parameters are set at their estimates. To reach the steady state, the model is simulated for a long enough "burn-in" period. We check across the paths from several independent simulations to make sure that they do converge to the same state.

### 7.1 Learning

What happens to a firm if it ignores the information given by the performance of past products? Can the whole industry do as well as before if it starts ignoring that information? Our first set of counterfactuals are designed to understand how much learning matters for the movie business. We first examine the case where a single firm stops learning, which is illustrated in the top row of Figure 7. The industry is at steady state at the beginning of the plotted period. Starting from the tenth year, firm 1 (corresponding to a major studio) treats the similarity among the products as zero, i.e., $\lambda = 0$.

There are several predictions. First, the firm invests in slightly fewer movies per year. Second, the firm shifts towards smaller-budget movies. This is because the absence of learning means that the firm faces a much larger uncertainty in the $\mu$'s of the arrivals, which makes it avoid big-budget movies. We also see a sizable decrease in the average latent quality. This is because without learning the firm is less effective in selecting better movies. The decreases in budget size and latent quality together suggest a decline in the industry profitability as well as the consumer welfare.

It is instructive to compare these predictions with those where the other firms stop learning as well, which are displayed in the middle row of Figure 7. The subjects of the plots are still the movies of firm 1, but we see much larger effects on budget size and latent quality, and an gradual *increase* in the number of movies produced. To understand the differences, notice that in the first counter-factual, the other firms are still selective about the latent qualities of the movies they produce. By imitating, or in the language of the model, attaching to these movies, the arrivals for firm 1 are able to maintain a reasonable level of latent quality. In this sense, the other firms in the industry are learning *for* firm 1. This explains why the average latent quality stays at a positive level in the first counter-factual but keeps decreasing towards zero in the second counter-factual. The same reason applies to the larger decrease in budget size in the



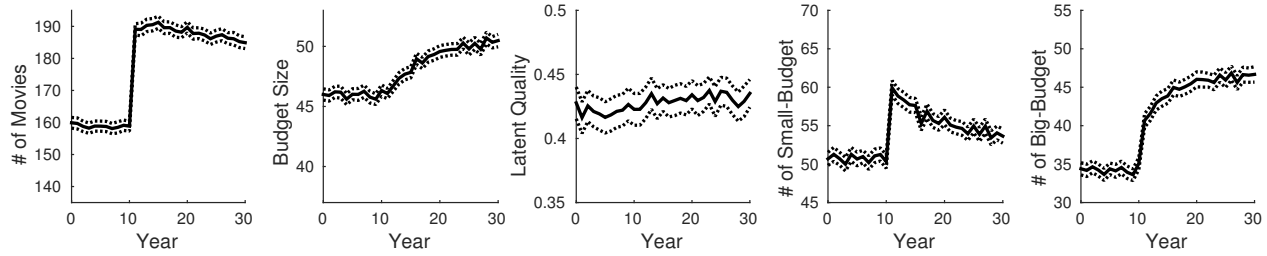

Each plot shows how the expectation of one of the following statistics change over time: (i) the number of movies produced by Firm 1 in each year, (ii) the average budget size of these movies, and (iii) the average latent quality of these movies. The expectations are further evaluated by multiple independently simulated paths. The dashed lines represent the 95% confidence band. In each path the simulation starts long before time 0 to reach steady state. For the top row, firm 1 stops learning (treats $\lambda$ as zero) after Year 10. For the middle row, all firms stop learning. For the bottom row, all the firms except firm 1 stop learning. Firm 1 corresponds to a major studio.

Figure 7: What If Firms Stop Learning

second counter-factual, which, combined with the mis-perceived originality in the arrivals, leads the firm to produce more movies per year.

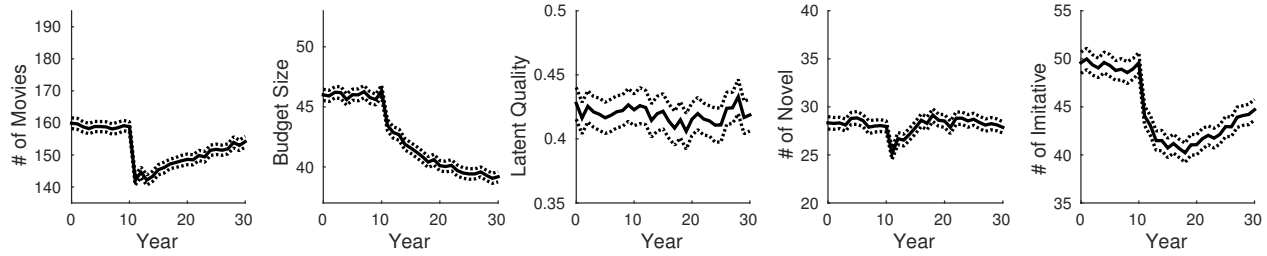
It is also instructive to compare with the case where all the other firm stop learning, but firm 1 does not. This is displayed in the bottom row of Figure 7. We still see decreases in the number of movies per year, average budget size and latent quality, but they are much more gradual compared with the first counterfactual. This is because the causes of the decreases are very different. Firm 1 does not see a larger uncertainty in the arrivals, nor does it become less selective about the latent quality. The budget size of its movies falls because the other firms switch to smaller movies; the latent quality of its movies falls because the the movies by other firms become of low quality. The eventual sizes of the decreases are no smaller, if not larger, than those in the first counterfactual. This suggests that the learning of the other firms is no less important to firm 1 than its own learning.[26]

[26]The findings echo the literature of learning-by-doing spillovers. See, for example, Irwin and Kelnow (1994), Benkard (2000) and Thornton and Thompson (2001).




The plots are generated in the same way as Figure 7. The two additional plots on the right are: (i) the number of movies with budget smaller than \$20m and (ii) the number of movies with budget bigger than \$70m. The arrival rate increases by 20% in Year 10 and stays at that level thereafter.

Figure 8: What Happens If More Arrivals




The plots are generated in the same way as Figure 7. The two additional plots on the right are: (i) the number of novel movies ($\leq$ 2 precursors), and (ii) the number of imitative movies ($\geq$ 10 precursors). The coefficient of risk aversion increases by 25% in Year 10 for all the firms and remains at that level thereafter.

Figure 9: What Happens If Firms Become More Risk Averse

### 7.2 Arrival Rate

When there is no existing products, then the first arrival is necessarily original. As more and more products are released, there is more room for imitation but less room for novelty. This reduces the uncertainty for the firms and allows them to accept bigger-budget movies. On the other hand, the decrease in novelty makes small-budget movies less appealing. Given this reasoning, the number of movies produced per year could affect the size of the movies. There are many other factors that could change budget sizes. To isolate the effects of the sheer number of movies, Figure 8 plots the counterfactual where we introduce a moderate increase, 20%, of just the arrival rate.

Consistent with our reasoning, we see not only a larger number of movies being produced each year, but also a larger average budget size. More details can be seen in the last two plots, where we look at the production of small-budget and large-budget movies separately. Initially, the production rate rises regardless of budget size. However, as the arrivals become less and less original, the production of small-budget movies gradually declines while the production of big-budget movies continues to grow.

The counterfactual suggests an explanation to the widely-acknowledged industry fact that movie budgets have been increasing over the past decades. In the data we see steady increases in the number of movies released per year in the late 1980s and early 1990s. Regardless of what the causes were, this increase of the production rate implies that the studios can have a better knowledge on consumer demand than they could many years ago. As a result, they are more willing to put money into a single bet. So this is really a "double expansion" of the industry: an increase in the number of products accompanied by an increase in the average size of each product.

### 7.3 Risk Aversion

Given the important role of risk aversion in our model, we now turn to examine what happens if there is a change in the firms' risk attitude. Changes in the level of risk aversion could be caused by factors such as the risk attitude of the studio managers (Lambert (1986)), the diversification of the parent company, or more broadly the condition of the financial markets. Figure 9 displays the scenario where the coefficient of risk aversion of all firms increases by 25% and stays at that level thereafter.

As expected, there are decreases in both the number and budget size of the movies produced each year: as the firms become more risk averse, they reject more arrivals. In particular, they reject a disproportionate number of big-budget movies, as they involve higher level of risk than their low-budget counterparts. However, what comes as a surprise is that there is a noticeable decrease in the number of imitative movies, while there is not much shift in the number of novel movies. In fact, the average number of precursors decreases. In other words, despite of firms

being more risk-averse, movies become more original on average.

To understand this rise of originality, we want to draw attention to several forces driving the degree of innovation or imitation. First, for any fixed size of budget, higher risk aversion implies that firms are less open to original movies. This force tends to decrease the level of originality and seems the most intuitive. However, there are two other less obvious forces working in the opposite direction. One comes from the fact that big-budget movies heavily rely on imitation (Figure 5). Thus as firms move to produce the smaller-budget movies, the degree of imitation may decrease. The other is that when a smaller number of movies are produced each year, there are fewer subjects for imitation, and the originality of the arrivals increases. Again, one can think of the extreme case that when there are no existing movies, the first production is necessarily original.

It is not difficult to imagine that if there is a *decrease* in the coefficient of risk aversion, the dynamics plotted in Figure 9 will be exactly reversed. In particular, the average originality will fall. This provides a possible explanation for the movie business model that relies more and more on "blockbusters – especially sequels and franchises."[27] Some people blame studios for being too focused on reducing risks to care for the originality of arts. Our counterfactual suggests, interestingly, that one perhaps should attribute the seemingly decline in originality to studios becoming *less* risk averse.

As a matter of fact, in 1989 and early 1990s, a series of conglomerate purchases and mergers that happened in the motion picture industry brought several studios new financial capabilities. In an attempt to reduce their risk exposure, studios started co-financing movies in the 1990s.[28] Both can be seen as factors that lower the level of risk aversion. In addition, decreasing risk aversion is consistent with the observation of industry experts that studios still offer "more balance than people think"[27] and keep producing "the kind of smartly-budgeted, star-driven [movie] that everyone claims never gets made anymore."[29] While a lower level of risk aversion allows the production of more movies and particularly the mega-budget, albeit imitative ones, it does not necessarily imply a reduction in the production of novel movies.

# 8 Concluding Remarks

By focusing on the U.S. motion picture industry, this paper studies new product entry in the presence of firm learning from the market performance of previous similar products. We make novel use of the Amazon and IMDb recommendation data to construct a similarity network amongst the products. The network allows us to examine the correlation in market performance between similar products and differentiate the levels of imitation across products. We introduce

[27] See "Are Blockbusters Destroying the Movies." *New York Times*, Jan 6, 2015.

[28] Co-financing is not explicitly modeled in this paper. Interested readers may look at Goettler and Leslie (2005).

[29] "Why Spielberg And Lucas Are Wrong About The Film Industry Implosion." June 20, 2013, *Forbes*.

and estimate an evolving network model that captures product entry over time. The model allows us to quantify the effects of learning and generate important insights into the balance of innovation vs. imitation. Given our findings, it is natural to ask how these findings generalize to other industries.

One finding is that one firm benefits substantially from the learning of other firms. The extent to which it holds in other industries depends on institutional details. In science, projects often seek inspirations and apply results from previous research. We even use the amount of imitation, namely the citation count, to measure how successful a paper is. Fashion design in America is well known for lacking intellectual property protection, and being forbearing about imitation. However, imitation renders a design obsolete very quickly, which likely limits the indirect benefits of learning (Pesendorfer (1995), Raustiala and Sprigman (2006)). In pharmaceutical industry, patents are essential for providing incentives for R&D, which limits how much one firm can benefit from the discoveries of others. Instead, these discoveries mean that the firm has to face more competitions.

We also find that big-budget products benefit more from imitation, but small-budget products favor novelty. This should be found in many other industries where uncertainty is high and firms are risk averse. For example, in software development, large applications (e.g., Windows, Office, Photoshop) are usually developed based on previous successful versions and started from prototypes, while millions of diverse small applications are developed and distributed on marketplaces like Mac App Store. In scientific research, the community faces the problem of funding allocation and the debate of small versus big science (Alberts (2012)). Our research demonstrates that a sheer increase in the number of products reduces uncertainty thus induces larger investment per project. In this sense, small projects serve as the guinea pigs for bigger projects.

Finally, we would like to address some limitations of our research. First, our model can be extended to weighted networks allowing for varying degrees of similarity between products. Such a network can be constructed with richer data on consumer co-watching behaviors and allows us to analyze imitation in greater detail. Second, it would be interesting to model the forward-looking behavior of firms. For one thing, forward looking means that firms will be more open towards innovation to explore alternative products of high demand. However, the size of the state space required to work with the network makes the problem very challenging; one probably has to start with approximate solutions. Third, while focusing on learning from previous products, our study ignores other sources of learning, most notably the market reception of related novels. It would be interesting to see how the successes of the products in adjacent industries lead to movie adaptations.

# 9 Appendix

## 9.1 A Toy Model

Consider a simple model of product entry with learning. There is one single firm and potential products arrive at a Poisson rate $\eta$. There is no production time, so if the arrival is accepted it gets released and generates revenue immediately. Now suppose that at time $t$ there arrives a potential product. Let us temporarily label this product by $j$. It is randomly assigned to be similar to one of the $n$ last released products: $j-1$, $j-2$, ..., $j-n$. The products older than $j-n$ become obsolete and are not imitated anymore. Let $y(j)$ denote the product that $j$ is similar to. The log return of $j$, $\mu_j$, is drawn from a normal distribution $\mathcal{N}(\lambda\mu_{y(j)}, \sigma^2)$, where $\lambda \in (0,1)$. At time $t$, the firm does not observe $\mu_j$ but knows $y(j)$ and $\mu_{y(j)}$, so that its expectation on $j$'s revenue is $\overline{\mu}_j \equiv \lambda\mu_{y(j)}$. Let $z_j$ be a product-specific cost shock known to the firm but not to us. We assume that $z_j \sim \mathcal{N}(0, \rho^2)$. The firm accepts $j$ iff $\overline{\mu}_j - z_j > 0$, and discards $j$ otherwise.

The model has five parameters: $\eta$, $n$, $\lambda$, $\sigma$ and $\rho$. The question is whether we can identify all of them. The answer is yes. Technically, the identification works as follows. Let $A$ be the set of accepted products within a period of length $T$. First, parameter $n$ can be simply identified with $\max_{j\in A} |j - y(j)|$. Next, noticing that $\xi_j \equiv \mu_j - \lambda\mu_{y(j)}$ is zero-mean normal with variance $\sigma^2$ and is i.i.d. across the accepted $j$, we can identify $\lambda$ and $\sigma$ by simply regressing $\mu_j$ on $\mu_{y(j)}$ for $j \in A$. Next, because a smaller $\rho$ makes the firm more selective in accepting products, the average expected log return of the accepted product, $\frac{1}{\#A}\sum_{j\in A}\overline{\mu}_j$, can be used to identify $\rho$. In the extreme case $\rho = +\infty$, there is no selection and the average should converges to 0. Finally, given $n$, $\lambda$, $\sigma$ and $\rho$, the production rate $\#A/T$ is strictly increasing in the arrival rate so it can be used to identify $\eta$.

When there are additional revenues or costs that are proportional to the ones used in calculating $\mu$, we can model them by adding an intercept parameter to the firm's decision. A product is accepted iff $\overline{\mu}_j - z_j - c > 0$. We want to ask if $c$ can be identified. From econometrician's perspective, the acceptance probability is

$$\begin{aligned} \Pr(j \text{ is accepted}) &= \Psi\left(\frac{\overline{\mu}_j - c}{\rho}\right) \\ &\simeq \Psi\left(-c/\rho\right) + \frac{\psi\left(-c/\rho\right)}{\rho} \cdot \overline{\mu}_j, \end{aligned}$$

where $\Psi$ ($\psi$) is the cdf (pdf) of the standard normal distribution. The second line is a linear approximation of the probit model around $\overline{\mu}_j = 0$. Now consider another set of parameters $(\eta', n', \lambda', \sigma', \rho', c')$ where $n' = n$, $\lambda' = \lambda$, $\sigma' = \sigma$, $c' = 0$ and

$$\rho' = \frac{\Psi(-c/\rho)}{\Psi(0)} \cdot \frac{\psi(0)}{\psi(-c/\rho)} \times \rho.$$

The acceptance probability becomes:

$$\begin{aligned}\Pr(j \text{ is accepted})' &\simeq \Psi(0) + \frac{\psi(0)}{\rho'} \cdot \overline{\mu}_j \\ &\simeq \frac{\Psi(0)}{\Psi(-c/\rho)} \times \Pr(j \text{ is accepted}).\end{aligned}$$

In other words, the acceptance probability is $\Phi(0)/\Phi(-c/\rho)$ times larger than before for *every* arrival product. If we choose the arrival rate

$$\eta' = \frac{\Psi(-c/\rho)}{\Psi(0)} \times \eta,$$

then the two sets of parameters are observationally equivalent. So had we specified a linear probability model instead of probit, parameter $c$ would not be identified.

### 9.2 Details on the Attachment Process

Fix a point of time $t$, the set of existing nodes and their network $Y$. The arriving node is $j$. Let $p_{k,j}$ be the 1st-stage attachment probability between $j$ and an existing node $k$. The probability that there will be no link between $j$ and $k$ after the two-stage attachment process is

$$1 - \Pr(y_{k,j} = 1|\mathcal{I}_t) = (1 - p_{k,j}) \cdot \prod_{\ell \sim k}(1 - p_{\ell,j}\omega),$$

where $\ell \sim k$ indicates that $\ell$ and $k$ are linked in $Y$.

In principle, given the value of $\Pr(y_{k,j} = 1|\mathcal{I}_t)$ for all $k$ (as specified by (2)), one could solve for the $p_{k,j}$'s. However, it poses a big computational burden to solve a nonlinear system with thousands of equations every time an arrival needs to be simulated. One heuristic approach is to seek approximate solutions by postulating that $p_{k,j} \simeq p_{\ell,j}$ for $k \sim \ell$. Given that the network features many layers of homophily (firm, release time, budget, latent quality), it does not seem an unreasonable assumption. In this case,

$$1 - \Pr(y_{k,j} = 1|\mathcal{I}_t) \simeq (1 - p_{k,j})(1 - p_{k,j}\omega)^{d_k(Y)},$$

where $d_k(Y)$ is the degree of $k$ in $Y$, i.e., the number of links connecting $k$. Taking the log of both sides, we have

$$\log\left[1 - \Pr(y_{k,j} = 1|\mathcal{I}_t)\right] \simeq \log(1 - p_{k,j}) + d_k(Y)\log(1 - p_{k,j}\omega).$$

Given that the attachment probabilities are generally small (less than 1%), we may use the

first-order Taylor approximation of log to obtain

$$-\Pr(y_{j,k}=1|\mathcal{I}_t) \simeq -p_{k,j} - d_k(Y)p_{k,j}\omega,$$

which implies

$$p_{k,j} \simeq \frac{\Pr(y_{jk}=1|\mathcal{I}_t)}{1+\omega d_k(Y)}. \tag{6}$$

The denominator evens out the 2nd stage's added attachment probability to nodes with higher degrees. We use (6) to readily compute the 1st-stage probabilities.

We can make a comparison with the alternative specification where one uses the right hand side of (2) directly as the 1st-stage probabilities. Such a specification implies that nodes with higher degrees are more likely to be attached to, similar to the concept of preferential attachment (Barabási and Albert (1999)). This leads to two undesirable features in our context. First, the probability of an original arrival (without precursors) is invariant to the density of the existing network. However, a sparse network implies diverse products, which should leave less room for innovation. Second, the model can become non-ergodic as a single product keeps being attached to over time. By connecting to the new entries, the product reinforces its probability of being attached to despite time discounting.

### 9.3 Details on Posterior Computation

For the expoisition of this subsection we will fix a time $t$. We use $R$ for the set of released movies: $\{k : r_k < t\}$, and $Q$ for the set of yet to be released movies: $\{k : a_k \leq t, r_k \geq t\}$. In our model the entire path up until $t$ consists of $\mathcal{I}_t$ and $\mu_Q$. It is not difficult to see that the probability of the entire path up until $t$ can be written as

$$\Pr\left(\mathcal{I}_t, \mu_Q\right) = \Psi(\mathcal{I}_t) \cdot \prod\nolimits_{k\in Q\cup R} \Pr\left(\mu_k|y_k, \mathcal{J}_{a_k}\right).$$

The product term includes the arrival probabilities of the latent qualities. $\Psi(\mathcal{I}_t)$ is the part that includes the probabilities of the Poisson arrivals, attachments, budget sizes and production decisions. Most importantly, all these do not involve the latent qualities of the yet to be released movies, hence $\Psi$ is a function of $\mathcal{I}_t$ only.

Given the specification in (3), the product term in the last equation is a joint density of the latent qualities that depends on the similarity network, start dates and release dates of the movies. Because these are included in $\mathcal{I}_t$, we can write

$$g\left(\mu_{Q\cup R}; \mathcal{I}_t\right) \equiv \prod\nolimits_{k:a_k\leq t} \Pr\left(\mu_k|y_k, \mathcal{J}_{a_k}\right).$$

Then by the definition of conditional density, we have

$$\begin{aligned}\Pr\left(\mu_Q|\mathcal{I}_t\right) &= \Pr\left(\mathcal{I}_t,\mu_Q\right)\cdot\left[\int\Pr\left(\mathcal{I}_t,\mu_Q\right)d\mu_Q\right]^{-1}\\ &= \Psi(\mathcal{I}_t)g\left(\mu_{Q\cup R};\mathcal{I}_t\right)\cdot\left[\Psi(\mathcal{I}_t)\int g\left(\mu_{Q\cup R};\mathcal{I}_t\right)d\mu_Q\right]^{-1}\\ &= g\left(\mu_Q|\mu_R;\mathcal{I}_t\right).\end{aligned}$$

Given the specification in (3), one representation of the unconditional density $g$ is

$$\mu_k=\sum\nolimits_{\ell:a_\ell\leq t}W_{k\ell}\mu_\ell+\upsilon_k.$$

where $\upsilon_k\sim\mathcal{N}(0,V_{kk})$. $W$ is a square matrix of the size $\#\{k:r_k<t\}$, and $V$ is a diagonal matrix of the same size. Their nonzero entries are:

$$\begin{aligned}W_{k\ell} &= \frac{\lambda\phi^{|r_k-r_\ell|}}{1+\lambda\sum_{k\sim\ell,a_k<a_\ell}\phi^{|r_k-r_\ell|}},\text{ if }\ell\sim k\text{ and }a_\ell<a_k,\\ V_{kk} &= \frac{\sigma^2}{1+\lambda\sum_{k\sim\ell,a_k<a_\ell}\phi^{|r_k-r_\ell|}}.\end{aligned}$$

In matrix form we can write in the matrix form

$$\mu_Q=W_{QR}\mu_R+W_{QQ}\mu_Q+\upsilon_Q,$$

or

$$\mu_Q=(I-W_{QQ})^{-1}(W_{QR}\mu_R+\upsilon_Q).$$

This tells us the distribution of $g\left(\mu_Q|\mu_R;\mathcal{I}_t\right)$. So

$$\Pr\left(\mu_Q|\mathcal{I}_t\right)=\mathcal{N}\left((I-W_{QQ})^{-1}W_{QR}\mu_R,\ (I-W_{QQ})^{-1}V_{QQ}(I-W_{QQ}')^{-1}\right)$$

To calculate the posterior on any subset $O\subseteq Q$ we can simply embark the calculation of the posterior on the entire $Q$. However, many times this is unnecessary and adds significant computational time in estimation or simulation because a large number of posteriors need to be calculated. In fact, $g$ belongs to the class of Gaussian Markov Random Field (Rue and Held (2005)), where two sets of nodes are conditionally independent given the values of a third set of nodes if the the third set separates the first two sets, i.e., every path connecting the two sets uses nodes in the third set.

By this result, one can show that the above equation still holds if we replace $Q$ with the collection of the nodes in $Q$ that are not separated from $O$ by $R$, and replace $R$ with the collection of the nodes in $R$ that are directly linked to some node in the replacement of $Q$. In the special case where $O$ is the single arrival movie $j$ and it is only linked to already released movies, the equation

Table 7: Monte Carlo Experiments for Supply Estimation

| Parameters | | **Percent** Bias | Percent Std. Dev. |
|---|---|---|---|
| Attachment | 2nd-stage Probability ($\omega$) | 0.1 | 1.2 |
| | Intercept ($\gamma_0$) | -0.2 | 0.7 |
| | Own Movies ($\gamma_1$) | 0.1 | 3.9 |
| | Time Difference ($\gamma_2$) | 0.5 | 2.6 |
| Obs. Characteristics | Budget Mean (No Precursors) ($\theta$) | -0.0 | 18.2 |
| | Budget Coeff. of Variation ($\chi$) | 2.4 | 5.3 |
| Coeff. of Risk Aversion ($\alpha$) | | 1.1 | 22.9 |
| Shock Size ($\rho$) | | -2.2 | 9.1 |
| Yearly Arrival Rate ($\eta$) | | 1.3 | 4.9 |

The model is simulated from 1995 to 2012 conditional on the data from 1975 to 1994. Parameters are set equal to their point estimates. Estimation is performed in the same way as on the real data, except that it treats demand parameter values as known. The experiment is repeated for 16 times. The first column shows the bias of the average estimate for each parameter, as percentage of the absolute value of the parameter. The second column shows the standard deviation of the estimates for each parameter, as percentage of the absolute value of the parameter.

reduces to (3), the arrival distribution of $j$.

### 9.4 Monte Carlo

We use Monte Carlo experiment to assess the supply-side estimator. The exercise consists of simulating the model under the parameter estimates to generate a dataset with the size similar to our real sample, and then applying the supply-side estimator to the dataset to recover the parameter values. We repeat this exercise a number of times to evaluate the distribution of the estimator. The results are displayed in Table 7. All the parameters are recovered with absolute bias smaller than 5%. The last column displays the dispersion of the estimator. The standard deviations are used as the parametric bootstrapping standard errors for the supply-side estimates (see Table 6).

# References


[1] Aguirregabiria, Victor and Chun-Yu Ho, 2011. "A Dynamic Oligopoly Game of the US Airline Industry: Estimation and Policy Experiments." *Journal of Econometrics*, Vol. 168(1).

[2] Alberts, Bruce, 2012, "The End of Small Sciences?" *Science*, Vol. 337, p. 1583.

[3] Andersen, Torben and Bent Sørensen, 1996. "GMM Estimation of a Stochastic Volatility Model: A Monte Carlo Study." *Journal of Business & Economic Statistics*, Vol. 14(3), pp. 328-352.

[4] Ainslie, Andrew, Xavier Drèze and Fred Zufryden, 2005. "Modeling Movie Life Cycles and Market Share." *Marketing Science*, Vol. 24 (3), 508-17.

[5] Barabási, Albert-László and Réka Albert, 1999. "Emergence of scaling in random networks." *Science*, Vol. 286, pp.509–512.

[6] Benkard, Lanier, 2000. "Learning and Forgetting: The Dynamics of Aircraft Production." *American Economic Review*, 90(4), 1034-1054.

[7] Bradlow, Eric, et al. 2005, "Spatial Models in Marketing." *Marketing Letters*, 16 (3-4), 267 - 278

[8] Chintagunta, Pradeep, 1994, "Heterogeneous Logit Model Implications for BrandPositioning." *Journal of Marketing Research*, Vol. 31, pp. 304-311.

[9] Ching, Andrew T., Tülin Erdem and Michael P. Keane, 2013. "Learning Models: An Assessment of Progress, Challenges, and New Developments." Marketing Science, Vol. 32(6).

[10] De Vany, Arthur and David Walls, 1996. "Bose-Einstein Dynamics and Adaptive Contracting in the Motion Picture Industry." *The Economic Journal*, Vol. 106, pp. 1493-1514

[11] De Vany, Arthur and David Walls, 1999, "Uncertainty in the movie industry: Does star power reduce the terrorof the box office?" *Journal of Cultural Economics*, Vol. 23: 285–318

[12] Dellarocas, Chris, Zsolt Katona, and William Rand, 2010. "Media, aggregators, and the link economy: Strategic hyperlink formation in content networks." Working Paper, NET Institute, New York.

[13] Einav, Liran, 2007. "Seasonality in the U.S. Motion Picture Industry." *RAND Journal of Economics*, 38(1).

[14] Einav, Liran, 2010. "Not All Rivals Look Alike: Estimating an Equilibrium Model of The Release Date Timing Game." *Economic Inquiry*, 48(2).

[15] Eizenberg, Alon, 2014. "Upstream Innovation and Product Variety in the U.S. Home PC Market." *Review of Economic Studies*, Vol.81, 1003–1045

[16] Elberse, Anita, and Jehoshua Eliashberg, 2003. "Demand and Supply Dynamics for Sequentially Released Products in International Markets: The Case of Motion Pictures." *Marketing Science*, Vol. 22(3).

[17] Eliashberg, Jehoshua, Anita Elberse and Mark Leenders, 2006, "The Motion Picture Industry: Critical Issues in Practice, Current Research, and New Research Directions." *Marketing Science*, Vol. 25(6).

[18] Eliashberg, Jehoshua, Sam Hui, and John Zhang, 2007. "From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts." *Management Science*, 53 (6), 881 - 893.

[19] Erdös, Paul and Alfréd Rényi, 1959. "On Random Graphs. I" *Publicationes Mathematicae*, 6:290-297.

[20] Elrod, Terry and Michael P. Keane, 1995, "A Factor-Anaiytic Probit Model for Representing the Market Structure in Panel Data." *Journal of Marketing Research*, Vol. 32, pp. 1-16.

[21] Goettler, Ronald and Phillip Leslie, 2005. "Cofinancing to Manage Risk in the Motion Picture Industry." *Journal of Economics & Management Strategy*, 14(2).

[22] Goettler, Ronald and Ron Shachar, 2001. "Spatial Competition in the Network Television Industry." *RAND Journal of Economics*, Vol. 32(4), 624-656.

[23] Han, Chirok and Peter C.B. Phillips, 2006. "GMM with Many Moment Conditions." *Econometrica*, Vol. 74(1), 147–192

[24] Hennig-Thurau, Thorsten, Mark B. Houston and Shrihari Sridhar, 2006, "Can good marketing carry a bad product? Evidence from the motion picture industry." *Marketing Letters*, 17: 205-219.

[25] Hitsch, Günter J., 2006. "An Empirical Model of Optimal Dynamic Product Launch and Exit Under Demand Uncertainty." Marketing Science, 25 (1).

[26] Holme, Petter and Beon Jun Kim, 2002. "Growing scale-free networks with tunable clustering." *Physical Review: E*, 65(2).

[27] Irwin, Douglas A. and Peter J. Klenow, 1994. "Learning-by-Doing Spillovers in the Semiconductor Industry." *Journal of Political Economy*, Vol. 102(6).

[28] Jackson, Matthew O., 2010. Chapter 5 in *Social and Economic Networks*. Princeton University Press.

[29] Jackson, Matthew O., and Brian Rogers, 2005. "Meeting Strangers and Friends of Friends: How Random Are Social Networks?" *American Economic Reivew*, 97(3): 890-915

[30] Lambert, Richard A., 1986, "Executive Effort and Selection of Risky Projects." *RAND Journal of Economics*, 17(1), 77–88.

[31] LeSage, James, 2008, "An Introduction to Spatial Econometrics." *Revue d'économie industrielle*, Vol. 123, p.19-44.

[32] Linden, Greg, Brent Smith and Jeremy York, 2003. "Amazon.com Recommendations: Item-to-Item Collaborative Filtering." *IEEE Internet Computing*, Vol 7(1), 76-80.

[33] Lu, Yingda, Kinshuk Jerath and Param Vir Singh, 2013. "The Emergence of Opinion Leaders in a Networked Online Community: A Dyadic Model with Time Dynamics and a Heuristic for Fast Estimation." *Management Science*, 59 (8), 1783 - 1799.

[34] Luo, Hong, 2014. "When to Sell Your Idea: Theory and Evidence from the Movie Industry." *Management Science*, 60 (12), 3067-3086.

[35] Mayzlin, Dina and Hema Yoganarasimhan, 2012. "Link to success: How blogs build an Audience by Promoting Rivals." *Management Science*, 58(9): 1651-1668.

[36] Mednick, Sarnoff, 1962, "The Associative Basis of The Creative Process." *Psychological Review*, Vol. 69 (3).

[37] Newman, Mark, 2003, "The Structure and Function of Complex Networks." *SIAM Review*, Vol. 45(2).

[38] Oestreicher-Singer, Gal, Barak Libai, Liron Sivan, Eyal Carmi, and Ohad Yassin, 2013. "Assessing Value in Product Networks." *Journal of Marketing*, 77:1-4.

[39] Oestreicher-Singer, Gal and Arun Sundararajan, 2012. "The Visible Hand? Demand Effects of Recommendation Networks in Electronic Markets." *Marketing Science*, 58: 1963–1981.

[40] Orbach, Barak and Liran Einav, 2007, "Uniform Prices for Differentiated Goods: The Case of the Movie-Theater Industry." *International Review of Law and Economics*, 27(2).

[41] Pesendorfer, Wolfgang, 1995, "Design Innovation and Fashion Cycles." *American Economic Review*, Vol. 85(4).

[42] Prag, Jay, and James Casavant, 1994, "An Empirical Study of the Determinants of Revenues and Marketing Expenditures in the Motion Picture Industry." *Journal of Cultural Economics*, Vol. 18(3).

[43] Raustiala, Kal and Christopher Jon Sprigman, 2006, "The Piracy Paradox: Innovation and Intellectual Property in Fashion Design." *Virginia Law Review*, Vol. 92, p. 1687.

[44] Ravid, S. Abraham, 1999. "Information, Blockbusters, and Stars: A Study of the Film Industry." *Journal of Business*, 72(4).

[45] Rue, Havard and Leonhard Held, 2005. Gaussian Markov Random Fields: Theory and Applications. CRC Press.

[46] Shriver, Scott, Harikesh S. Nair and Reto Hofstetter, 2013, "Social Ties and User-Generated Content: Evidence from an Online Social Network." *Management Science*, 59 (6), 1425-1443.

[47] Shen, Qiaowei and Ping Xiao, 2014a. "McDonald's and KFC in China: Competitors or Companion." *Marketing Science*, 33(2).

[48] Shen, Qiaowei and Hongju Liu, 2014b, "Demand Uncertainty, Dynamic Learning and Exit in Competitive Markets." Working Paper, University of Pennsylvania.

[49] Squire, Jason, 2005. The Movie Business Book, 3rd Edition, Simon and Schuster.

[50] Toivanen, Otto. and Michael Waterson, 2005. "Market Structure and Entry: Where's the Beef?" *RAND Journal of Economics*, 36, 680-699.

[51] Thornton, Rebecca A. and Peter Thompson, 2001. "Learning from Experience and Learning from Others: An Exploration of Learning and Spillovers in Wartime Shipbuilding." *American Economic Reivew*, Vol. 91(5).

[52] Uzzi, Brian, Satyam Mukherjee, Michael Stringer, and Ben Jones, 2013. "Atypical Combinations and Scientific Impact." *Science*, Vol.342: 468-472.

[53] Wallace, Timothy, Alan Seigerman, and Morris Holbrook, 1993. "The Role of Actors and Actresses in the Success of Films: How Much Is a Movie Star Worth?" *Journal of Cultural Economics*, 17(1), 1-27.

[54] Watts, Duncan J. and Steven Strogatz H. 1998. "Collective Dynamics of 'Small-World' Networks." *Nature*, 393 (6684).

[55] Weitzman, Martin L., 1998. "Recombinant Growth." *Quarterly Journal of Economics*, Vol.113, 331-360.

[56] Yang, Nathan, 2014, "March of the Chains: Herding in Restaurant Locations." Working Paper, Yale University.