October 22, 2018

Social Network Analytics, Empirical Exercise #4
Due on Monday, November 12, 2017 at 9:00am

Creative innovation through collaborations in feature film production

Setting up film production data: This exercise analyzes production companies of feature films and the films that they make. Filmmakers are in the fascinating position in which, unlike some other kinds of firms, they must constantly release novel, creative productions that are finely tuned into uncertain and ever-changing audience interests. Many organizations struggle to adapt to changing environments and audience interests, and struggle to remain competitive—consider the position of a firm IBM today, as opposed to twenty years ago. Yet, the same production companies that were dominant decades ago remain some of the most viable today. How are they able to adapt so effectively to new filmmaking technologies and audience interests? Production companies' collaboration patterns can explain some of this puzzle.

- The file "producers_and_films.csv" contains information on film production companies and the film projects that they produce.
  - Producers and films are listed by name, in the columns "prod_company" and "project", respectively.
  - There are also unique identifier keys for producers and films in the columns "pcindex" and "pindex", respectively. The keys require less memory usage than the full text strings in the names, so using the keys might make some of the computations run more quickly.
  - The country of each producer is also listed—for all of the analyses, only consider US producers so that we can focus on the Hollywood collaboration network more specifically.
  - There is also a column indicating the year that the film was released.
- The file "box_office_revenues.csv" contains information about the box office revenues for each film.
  - The box office revenue in dollars is given in the column "total_box".
  - Budget information, where available, is also given in the budget column. Values of zero are missing data.
  - There is also information about the number of screens the film was shown on in theaters. This is a good proxy for the missing budget data, because it captures the investment into film distribution and marketing made by the producer for each film. The column "release_coverage" is proportion that represents the number of theater screens a film was shown on, divided by the total number of screens available in the United States that year.
- The file "film_keywords.csv" contains a list of keywords used in each film. Keywords represent low-level plot, aesthetic, and stylistic elements of each film. A film can have up to a few hundred keywords that describe it. The keywords are useful because they help us keep track of new kinds of film features as producers create them. They also allow us to compare the similarity of films and their producers based on the film feature keywords they use in the films that they make.
- When analyzing production collaborations, we also want to make sure that we can take into account whether a production company is a subsidiary of another production company or not. The file "production_subsidiaries.csv" indicates the production companies that are subsidiaries of another production company in the data, and the first and last years—for cases where the subsidiary was sold or became an independent producer—the company operated as a subsidiary.

1. First, we want to know if filmmakers that engage in collaborations with one another are more innovative or not. We can measure innovation through the number of new, never-before-seen keywords that are used in a film. We can also measure innovation through the number of new combinations of existing keywords that are used in a film. To account for the natural

time cycle of the production process, consider a keyword or combination to be "new" if it has been introduced within the last three years.

We also want to know what kinds of collaborations contribute to innovation: are collaborations between large, "generalist" production companies more innovative? Or, are collaborations between large producers and more specialized, smaller producers more innovative? For all of the questions, consider a film producer to be a generalist if it is in the top quartile of the number of films released by producers that year. In general, a producer will be classified as a generalist if it makes more than one film in a year.

(A) Classify each film by the type of collaboration that it represents. There should be five types:

i. Peripheral solo productions: films made by a single specialist
ii. Central solo productions: films made by a single generalist
iii. Central co-productions: films made by a group of multiple generalists
iv. Peripheral co-productions: films made by a group of multiple specialists
v. Hybrid co-productions: films made by a group of generalists and specialists

Create a figure that illustrates the number of new keywords and new combinations of existing keywords that are introduced per type of film over the course of the data. On the x-axis should be years, and on the y-axis should be the count of new keywords or new combinations.

(B) Estimate one regression predicting the number of new keywords and another regression predicting the number of new combinations of existing keywords producers introduce in a year. Use as predictors the number of films a producer makes that year that year that fall into each of the three co-production types. So, there will be three collaboration predictors:

i. Central co-productions: number of Central co-productions a producer made that year
ii. Peripheral co-productions: number of Peripheral co-productions a producer made that year
iii. Hybrid co-productions: number of Hybrid co-productions a producer made that year

Also include control variables for a producer's box office revenue that year, how many years the producer has been in operation, whether or not the producer is a subsidiary, and a time trend for each year.

Since it possible for some types of films to be more innovative than others, also control for the content of producers' films. To do this, perform a multidimensional scaling using two dimensions that uses as the input the Jaccard distance between each producer based on the co-occurrence—the overlap—of keywords that they use in their films. To account for the natural time cycle of the production process, use as the comparison set for similarity the current year as well as the two years before the current year. You can calculate Jaccard distance using the `dist()` command from the proxy package, and you can perform the multidimensional scaling using the `cmdscale()` command from the stats package, which is automatically loaded when R starts. Use the two coordinates produced by the multidimensional scaling as controls in the regression.

Similar to the political parties on the previous exercise, the outcome variable is a count, so we can use a regression adopted for data of this form using the MASS package, using a model specified in the form of

```
glm.nb(new keywords variable ~ Central co-productions + Peripheral co-productions
+ Hybrid co-productions + Coordinate 1 + Coordinate 2 + Total box office +
Number of years in operation + Is subsidiary + factor(year), data, offset(total
films made that year, for which there is keyword information))
```

The offset accounts for the fact that making more films provides more opportunities to produce new keywords—it allows us to estimate the outcome as a per-film rate.

Are collaborations related to introducing more new keywords and more new combinations of existing keywords? What type of collaboration seems to result in the most new keywords and new combinations?

2. What might explain how some collaborations result in more innovative films than others? It could be that when producers collaborate with other producers that are too similar to themselves, their experience is less diverse and it is more difficult to come up with new innovations. We can measure the extent to which a producer collaborates with similar producers as the average Jaccard distance between a producer and the other producers it works with based on the co-occurrence of keywords the producers use.

   Generate this measure yearly for each producer—again, to account for the natural time cycle of the production process, use as the comparison set for similarity the current year as well as the two years before the current year.

   Create one figure that illustrates how the distance between a producer and the other producers it works with relates to the number of new keywords a producer introduces each year, and another for the relationship between this distance and the new combinations of existing keywords each year. A useful way to do this is by using a "loess" smoother that plots a flexible trend line that illustrates the level of a variable on the y-axis at different levels of a variable on the x-axis. Loess stands for "locally estimated scatterplot smoothing"—it fits a locally-weighted regression line over the underlying scatterplot, so it provides a tool to observe nonlinear relationships between the two variables.

   You can set up a loess plot using the ggplot2 package and running a command of the form

   ```
   ggplot(data, aes(average Jaccard distance, new keywords)) + geom_smooth(method
   = "loess", se = T) + labs(x = "Average Jaccard distance", y = "New keywords"))
   ```

   You can also export the plot quickly to pdf using ggplot2's functionality

   ```
   ggsave("loess_new_keywords.pdf", width = 7, height = 7, units = "in")
   ```

   which lets you control the size of the pdf that is saved.

   What does the pattern suggest about what kinds of collaborative partnerships might result in more creative innovation? Does this help to explain the results from Question 1?

3. Next, let's analyze whether collaborations influence a production company's financial returns. Since the budget information is so sparse, we will use the theater screens release coverage as a proxy for how much producers spend on each film that they make. Define each producer's yearly return as its yearly box office revenue divided by the total release coverage it invested in for that year for its films.

   To be able to make comparisons more equally across the years of the data, we'll normalize each producer's box office return compared to the returns that all producers earned that year. To do this, subtract the mean return of all producers for that year from a producer's individual return and divide it by the standard deviation of the returns for all producers that year:

   $$standardized\ return_{it} = \frac{return_{it} - mean\ return_t}{standard\ deviation\ return_t}$$

   for each producer i in each year t.

   Estimate a regression predicting producers' standardized return. Use the same predictors and controls as in Question 1. Since the outcome is not a count, you can estimate this model with

   ```
   lm(standardized return ~ Central co-productions + Peripheral co-productions +
   Hybrid co-productions + Coordinate 1 + Coordinate 2 + Total box office + Number
   of years in operation + Is subsidiary + factor(year), data)
   ```

   What do the results suggest?

4. Collaborations can be financially risky because of the coordination required to integrate multiple producers' experiences into a making new film. Do producers gain anything from these collaborations creatively or financially in the long term?

   (A) Estimate one regression predicting the count of new keywords introduced in a producer's *solo* produced films and another predicting the count of new combinations of existing keywords introduced in a producer's *solo* produced films in a year.

In the first regression, use as a predictor the cumulative number of new keywords a producer has introduced in all of its films through the current year that were made in collaborations. In the second regression, use as a predictor the cumulative number of new combinations of existing keywords a producer has introduced in all of its films through the current year that were made in collaborations. Use the same set of controls as in Questions 1 and 2. The outcome is a count, so use `glm.nb()`.

Does creative innovation gained through collaborations make a producer's solo-produced films more innovative? What does this suggest?

(B) Accounting for a producer's engaging in collaborations, does introducing new keywords and new combinations of existing keywords result in higher box office returns?

To gain insight into this, estimate the same form of regression from Question 2, but in one regression add in a predictor for the number of new keywords introduced, and in another regression add in a predictor for the number of new combinations of existing keywords introduced.

Does this result help explain why producers might engage in collaborations, even though they can be financially risky?