November 17, 2021

Social Network Analytics, Empirical Exercise #5

Due on Friday, December 3, 2021 at 11:59pm

Diversification and social status in venture capital

Setting up the status hierarchy: Earlier in the course, we looked at how the co-investment networks of venture capital firms influenced different aspects of their performance. In this exercise, we will analyze how the co-investment networks of venture capital firms influence their strategies, in terms of the types of startup companies that they invest in. Investors are constantly trying to find the next, best project to work on, but how to diversify into new industries is often unclear. Diversification is challenging because it requires new skills and expertise. Firms with high social status may be more effective at managing firms in disparate industries because they are able to leverage their position in order to get more recognition for their ventures, even if they require distinct sets of skills.

- The file "investors_and_deals.csv" contains information on investment firms and the deals that they invest in.
  - Investors are listed by a unique identifier
  - Deals that investors participate in are listed by another unique key
  - The column "Lead_investor" indicates whether a firm was a lead investor on a deal
- The file "companies_details.csv" contains information about the industry category the startup companies that are invested in by venture capital firms are in, and whether the startup is earning revenue or not.
- The file "deal_details.csv" contains information about the particular deal. It relates companies to deals using a unique key for each company, which maps back to the investors and deals file using the the consistent key for each deal. This file also indicates the date and Deal Type of each deal.
- Last, the the file "investor_details.csv" contains information about the firms that invest in each deal. This includes their name and Investor Type, and can be mapped back to the investors and deals file using the consistent key for each investor.

1. First, we want to know if higher-status firms are more likely to diversify their investments into different industries than are lower-status firms or middle-status firms. Since we want to make comparisons among firms within the same status hierarchy, only consider in the analysis investors of the Venture Capital type, given in the variable "Investor_Type" in the file "investor_details.csv". Similarly, to capture the period of investing where there is a well-defined status hierarchy, consider deals that have occurred from 1990 onward.

   We can define a status relationship for a pair of firms (A→B) as the proportion of times that Firm A has served as a lead investor in deals it has participated in with Firm B:
   $$\frac{count(deals\ with\ B\ in\ which\ A\ is\ lead)}{total\ deals\ with\ B}$$
   Let this proportion be the the entries of a matrix representing a relationship between each of the investors. The diagonals of this matrix should be 0. Then, each investor's status can be represented as the Bonacich centrality of this matrix—an investor's status is represented by its ability to be a leader on deals, as well as to be connected to other firms that lead their own deals. You can calculate Bonacich centrality using `power_centrality()` with the argument `exponent = 0.75` to represent the commonly-used beta parameter.

   For the analysis, only consider firms that are actually a part of the status hierarchy, i.e., have co-invested with other firms and have non-missing values for status. To allow older ties to weaken over time, you can exclude ties that have not been renewed after five years.

   (A) Traditionally, venture capital analysis have only considered the concentration of a venture capital firms investments into different portfolio categories. The more concentrated a firm's investments, the less diversified it is.

In Exercise 3, we used a Herfindahl index to measure political concentration, following the formula

$$Herfindahl = \sum_{i}^{n}(market\ share_i)^2$$

Create a variable that measures the cumulative concentration of each venture capital firm's portfolio through each year it has made an investment. You do not need to calculate the concentration manually—you can use the `hhi` package to set up the variable for each firm for each year. Use as the industry category the company-level variable "Primary Industry Code"

It is possible that diversification may increase **linearly** as we move up the status hierarchy, or the pattern may be **nonlinear**, representing a higher-order polynomial relationship such as a parabola.

In order to test a simple parabolic relationship for a predictor variable, we can include a squared term in the regression model. This is the same as interacting the predictor variable with itself. As a result, these models can take the form in R of

```
outcome variable ~ predictor variable + I(predictor variable^2)}
```

Run a regression predicting a venture capital firm's concentration in a year based on its status in the prior year it made an investment and also include the square of this term.

One motivation that might influence venture capital investors to diversify might be to minimize risk. In order to isolate the effect of status from this alternative rationale, include lagged control variables for a venture capital firm's risk exposure:

- whether a venture capital firm tends to originate its own deals: for more than 50% of the companies it invests in, it invests in the first investment round this company has received
- whether a venture capital firm tends to invest in the IT sector: more than 50% of the companies it invests in are in the company-level variable Primary Industry Sector "Information Technology"
- whether a venture capital firm tends to invest in early-stage startups: more than 50% of the companies it invests in are of the Deal Type 1 "Early Stage VC", "Accelerator/Incubator", "Seed Round", or "Angel (individual)"

Also include controls for a firm's age in years and also for the year as a linear control.

Similar to district-fixed effects models in Exercise 3, to incorporate venture capital firm-fixed effects estimate the model using `plm` with `model = "within"` and `effect = "individual"`.

What is the relationship between status and diversification?

(B) One downside to traditional measures focusing on concentration is that they ignore how different industry categories might be related. For example, diversifying into Application Software from Entertainment Software is more likely to share the same set of knowledge and skills than diversifying into Cruise Lines from Commodity Chemicals.

New approaches to tackling this question are able to consider the relatedness of industry categories using network analysis. Create a new measure of diversification that takes the relatedness of industry categories into account. First compute the relatedness of each industry category as the Jaccard distance between each pair of industry categories for each year, using the company-level variable "Primary Industry Code". Base the similarity on the co-occurrence of industry categories in investors' portfolios cumulative to the current year.

Then, relate these distances to each venture capital firm by summing the distances between each pair of industries that appear in its portfolio cumulatively through each year that it makes an investment. Take a new measure of diversification as its *niche width*:

$$niche\ width = 1 - \frac{1}{1 + \frac{\sum_{it} distances}{total\ number\ of\ industries - 1}}$$

In this way, diversification is represented as an "average distance" between each of the industry categories that appear in a firm's portfolio in each year. If an investor only invests in a single industry category, Run the same regression as in 1A using this new variable. Since the outcome varies from 0 to 1, estimate the model using `glm` with `family = quasibinomial(link = "logit")`. The approach for incorporating fixed effects for this model is different: include in the model the average values for all of the predictors, except for the year, for each firm over its lifetime. This follows the approach of a "Mundlak" estimator for venture capital firm heterogeneity.

What is the relationship between status and diversification?

(C) Next, let's check the shape of the regression curve to get a sense of the parabolic curvature. First, re-run the regression from 1B just using lagged status and the status squared term and not using any of the additional controls. Store the results from this regression in an object. Next, set up a data object with a range of values of the lagged status variable—100 values ranging from the minimum to the maximum of this variable will be sufficient. Use the `predict()` function to generate fitted values for each of these status values from the regression. Generate 95% confidence intervals for the fitted values by multiplying the standard error of the fit by ± 1.96.

Set up a plot with the fitted values and their confidence intervals across the range of lagged status, which should be on the x-axis. What does the curve suggest about the diversification strategies of low, middle, and high-status venture capital firms?

2. Which venture capital firms are more effective at diversifying their portfolios? In a regression, it is possible to determine how two variables jointly predict an outcome by specifying an "interaction" term. This term can be included in the model by including a term that multiplies the variables together—`variable1:variable2`—in the formula of the model. If this term is positive, it indicates that there is a synergistic effect of the two variables—high levels of both together have a positive effect on the outcome variable.

(A) Investments for venture capital firms are successful when they generate cash for the venture capital firm. Consider a count of successful investments venture capital deals that generate cash: startups that become publicly traded or are acquired for profit. This count should be the cumulative number of deals for a venture capital firm that fall into the Deal Type 1 categorization "IPO", "Merger/Acquisition", or "Buyout/LBO".

Run a appropriate regression, considering the form of the outcome variable and incorporating venture capital firm fixed effects, predicting the number of successful investments as a function of lagged status, lagged diversification, and interaction of lagged status and lagged diversification. Use the niche width measure of diversification and include the same controls from the regressions from 1A and 1B.

Is this interaction related to having more successful investments?

(B) Similar to 1C, we can use a visualization to better understand the relationship between the variables in the regression. We can accomplish this using a 3d scatterplot or a contour plot generated from the fitted values of the model.

Re-run a similar model from 2A with just lagged status and lagged diversification and without using firm fixed effects, e.g., using `glm()` with `family = "poisson"`, and assign it to an object. Next, generate a range of values for lagged status and lagged diversification, similar to 1C. Use the function `expand.grid()` to range of combinations of status and diversification, and then use predict to get the fitted values for each combination of diversification and status. Below is some code that will generate a 3d scatterplot.

```
# regular 3d plot
scatter3D(values$diversification, values$status, values$successful_investments)

# interactive 3d plot
plot3d(values$diversification, values$status, values$successful_investments)
```

The command `scatter3D` is from the package `plot3D`, and the command `plot3d` is from the package `rgl`.

A contour plot can be executed in a similar manner using the following code, from the `plotly` package.

```
p1 = plot_ly(
values,
x = ~status,
y = ~niche width,
z = ~fit,
type = "contour",
autocontour = FALSE,
contours = list(
end = max(values$fit, na.rm = TRUE),
size = abs(max(values$fit,  na.rm = TRUE) - min(values$fit,  na.rm = TRUE))/20,
start = min(values$fit,  na.rm = TRUE),
showlines = FALSE
),
line = list(smoothing = 0.85),

colorscale = "Greys"
) %>%
layout(font = cmodern) %>%
colorbar(len = 1, nticks = 10, title = "Estimated successful \n investments") %>%
    layout(yaxis = list(title = "Niche width")) %>%
    layout(xaxis = list(title = "Status"))
```

What do the patterns suggest about which venture capital firms are most or least successful overall at diversifying their portfolios?

3. The parabolic relationships from 1B and 1C suggest that low and high-status venture capital firms may share similar tendencies to diversify, but the estimates from 2A suggest that high-status firms are better at diversifying. Why might this be the case?

   Next, we will examine what strategies might make high-status firms better at diversifying the portfolios. One way high-status firms might diversify more effectively is that while their own expertise might be farther away from an industry category, they can use their social influence on deals for which they are the lead investor to coordinate the assistance of co-investor syndicate partners with expertise close to the industry category.

   Use a multidimensional scaling of two dimensions to determine the position of each venture capital firm's investment portfolio based on its cumulative investments up through each year. Each investor should have its coordinated updated each year. Use as the input to the scaling the Jaccard distance between each producer in each year, based on the industry categories in their portfolio given by "Primary Industry Sector".

   Similar to the partitioning around medoids clustering, define a medoid for each industry category as the coordinates represented by a venture capital firm that only invests in that category in a particular year. If no firms invest exclusively in the category, you can use as the medoid the firm with the most investments in this category.

   For each deal, define the distance between a firm's experience and the industry category as the Euclidean distance between the firm's coordinates and the coordinates of the medoid for the industry category. This distance can be found with the `dist()` function from `proxy` or `stats`.

   (A) Run a appropriate regression, considering the form of the outcome variable and incorporating venture capital firm fixed effects, predicting the average distance between a firm's syndicate partners and the industry category medoids for the deals that it invests in in a given year, as a function of a firm's lagged status, the firm's *own* average distance from the industry category medoids for the deals that it invests in in a given year, and the interaction between these two variables. To measure the coordination process a firm engages in when it is a lead investor, only include in the variable calculations the deals for which a firm is the lead investor.

   Include the same set of controls as the regressions from 1A, 1B, and 2A.

What does the regression suggest about how high-status firms might use their influence to coordinate other firms' expertise on deals that are further away from their own expertise?

(B) Set up a 3d scatterplot or a contour plot similar to 2B illustrating the relationship between status, a firm's own distance from the industry categories that it invests in, and the fitted values from the regression.

What do the patterns suggest about how high-status venture capital firms develop strategies to diversify more effectively?

Extra credit (2 points)

Counterfactuals of venture capital investment.

One issue with determining the likelihood of venture capital investment is constructing a reasonable counterfatual sample. That is, our prior analysis is based upon prior investments that have been realized, but to estimate the venture capital investor's decision most accurately we want to also estimate their decision in the context of all potential deals they might have pursued at the time they chose to make an investment.

To estimate investments among these potential deals, we need to construct a counterfactual sample of deals a venture capital firm could have chosen from at any time they chose to make a deal. To construct the counterfactual sample, create a set, for each deal that a venture capital firm invested in, of all deals that occurred within a 30-day window of the date of this actual deal that occurred within that deal's industry sector. Use a logistic regression to examine whether high-status firms tend to invest in deals that are further away from their own industry expertise.

The regression can take the form

```
glm(realized_investment ~ lag_status:lag_distance_from_own_expertise +
↪  lag_status_sqaured:lag_distance_from_own_expertise + lag_controls, data =
↪  data, family = binomial(link = "logit"))
```

Where `realized_investment` takes on a value of 1 for actual investments and 0 for investments in the matched sample that have not occurred (but were made in the same industry sector within 30 days). Positive coefficients on the interaction variables for `lag_status:lag_distance_from_own_expertise` and `lag_status_squared:lag_distance_from_own_expertise` should indicate that higher-status venture capital firms are more likely to make investments in more diversified industries. Do the results from the counterfactual analysis align with the results from Question 1?