**I. Introduction**

Investor Networks in Venture Capital:

with Silicon Valley Bank

Trevor Gower

M.S. Data Science: GalvanizeU, University of New Haven

44 Tehama St. San Francisco, CA

[trevorgowerr@gmail.com](mailto:trevorgowerr@gmail.com)

Venture capital is a notoriously difficult field for investors. Companies are young, with uncertain financials at best, seeking to break into emerging markets. A successful investment is as much about understanding the people behind the company as it is in understanding the market itself.

Relationships and personal trust play a major role in determining which companies investors will participate in. Similarly, lending institutions develop ongoing relationships with investment firms and with particular partners in those firms. These relationships run between venture partners, firm founders, and the lending officers; in this case we will focus on the investor relationships.

Silicon Valley Bank is the leading institution of venture lending. It has strong ties with the top venture capital firms, particularly in the Bay Area, and is known for providing debt financing to promising companies during early venture stages. Early venture stages range from the seed round to Series A through C, and Silicon Valley Bank often becomes involved at the Series A stage.

Investor social networks have been found to influence investment patterns.This is true not only due to shared performance in past deals, but also human social structures apart from financial considerations.

Silicon Valley Bank, hereafter to be referred to as SVB, was interested in exploring various uses of an investor network graph. Information about the relations of investors to each other and the relationship of investors to successful deals with the bank, could hold valuable insights.

PitchBook is a website devoted to covering venture capital, from seed stage to public offerings. In particular, it is known for keeping records of individual people with firms, including partners in venture capital firms and high-ranking employees and found of startups. Data from PitchBook, combined with SVB’s internal resources, presents a valuable source of potential insights.

**II. Overview**

**A. Theory of the Solution**

This project’s first goal was to establish a graphical model of the data in R and develop descriptive tools for nodes.

Data from the Pitchbook website is converted into a graphical model, with individual investors as nodes. Edges are directional and indicate that there has been at least one time when the head (or “to”) node followed the tail (or “from”) node. Weights indicate the number of times the “to” investor has followed the “from” investor.

With the graph formed, centrality, path length, and dense subgraphs were investigated. All of which may be used in business decision making or further machine learning.

Centrality is a common indicator of the relative importance of a node. Path length between nodes can be used as a proxy for the nearness of a relationship. Dense subgraphs represent clustering opportunities.

This project’s second goal was to begin implementation of exponential random graph models on investor social networks.

ERGM’s enable the analysis and prediction of new social behaviors.

ERGM’s have the unusual distinction of training directly on the graph shape. If a node is part of three edge, and two of those edges are part of a triangle formation, these can be trained on as features: 3 edges and 1 triangle.

An ERGM model estimates the posterior that any given node will form an edge with another random node. It does this by treating the structural features as homogenous across the graph - being a part of a social triangle is assumed to have the same general effect.

The posterior distribution yields the probability that we would see a graph matching our current graph given a set of parameters.

P(Y = y|θ) = exp(θ’ f(y))/c

* Where y is the binary adjacency matri
* f(y) consists of the structural features
* θ are the parameters to be learned
* c is the normalizing constant.

Modeling performance can be improved by the inclusion of other factors such as weights on existing edges, distances between nodes, and nodal covariates - node features that can delineate separate distributions.

ERGM models can be developed both numerically and analytically, as is performed in this project. In the analytic case, the max likelihood estimate is solved for.

**III. Implementation**

**A. Choice of Tools**

Python and R were used for this project. Prior work in Python with the Pitchbook data made it the best choice for preparing the data prior to graph formation.

R has a large suite of graph packages, including Graph, RGraphviz, RGBL, BNLearn, etc. It was therefore selected for both the native support of the graph structure and for the variety of machine learning packages available.

Graph exploration and centrality measurements were done with Graph, RGBL, and RGraphViz, the latter of which were developed by Bioclite. ERGM modeling was performed using Statnet.

**B. Choice of Data**

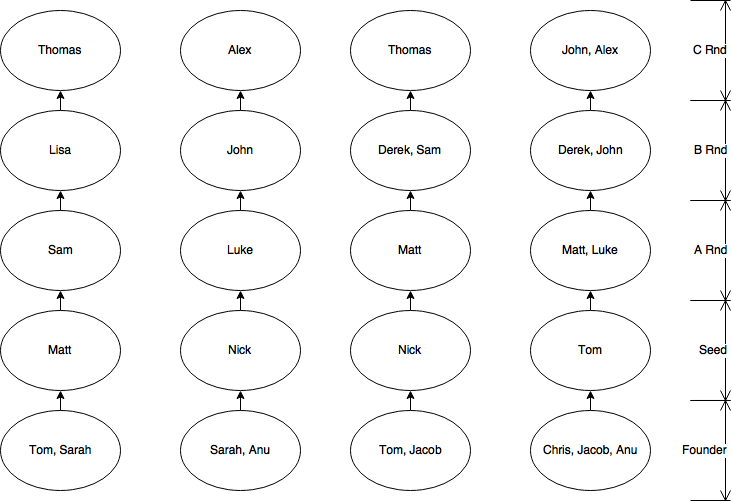
Pitchbook, as previously stated, contains considerable information regarding the involvement of individual people in companies and investments.

Companies of interest were selected and deals ranging from Angel and Seed to Series C were downloaded.

In total, nearly 4,000 investors with over 27,000 unique directional edges were extracted from roughly 2,000 deals.

Pitchbook’s Excel Download tool was used to download the information for the A deals for these companies and then to create CSV files. This data included the date of the deal and information identifying the company, the CEO of the company, and the Investors that took part in the deal. In Pitchbook’s case, an Investor refers to a specific venture capital firm, which may or may not have an individual person (Partner), attached to the deal.

**Conceptual Chart of Input Data**



Examples of interpreting the chart:

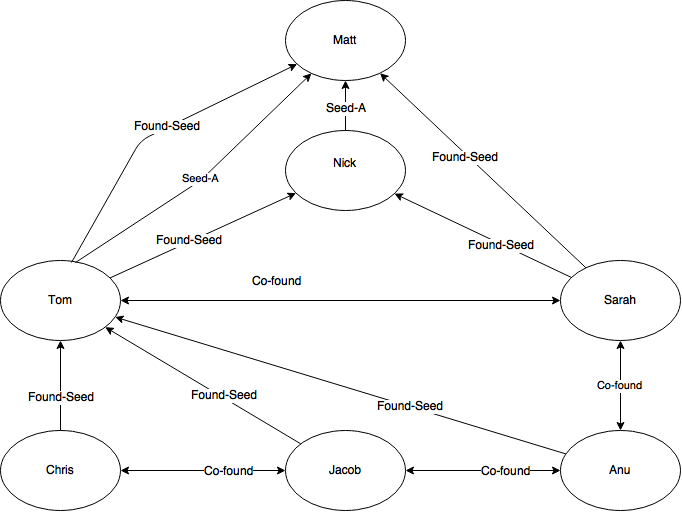
Tom and Sarah founded the left-most company, and Matt invested in the Seed Round of that company.

Derek and John both invested in the B Round of the right-most company, after Matt and Luke invested in the A round of that company.

Possible Questions:

If Sarah founded a new company, what is the probability that Tom would invest in the seed round? What is the probability that Matt would invest in either the Seed or the A round?

What if Sarah founded a new company along with Chris? How would that affect the previous probabilities? What is the probability that Thomas would invest in the C Round?

****

**Extracting First-Order Relationships**

**IV. Findings**

**A. Centrality and Descriptive Statistics**

The relationships extracted from the Pitchbook data were fed into an ftM2graphNEL, a graph sub-class with directed edges.

*inv\_graph = ftM2graphNEL(data.matrix(rel\_data), data.matrix(w\_data))*

*A graphNEL graph with directed edges*

*Number of Nodes = 3393*

*Number of Edges = 27030*

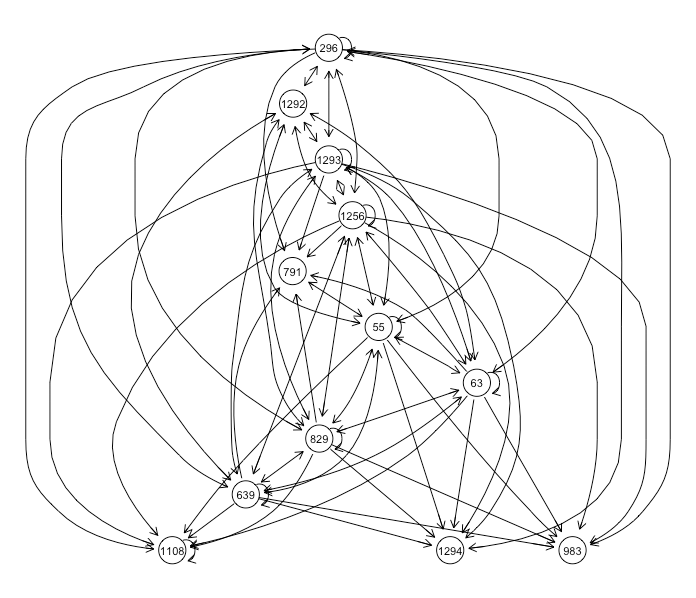
A graph with 3393 nodes has 11,509,056 possible edges, not counting any self-loops.

Therefore, despite the presence of over 27,000 edges, we have a very sparse data set. Only .23% of possible edges are populated in the graph.

The RGBL package contains various means by which to compute the shortest path between any two nodes. The “johnson.all.pairs” function was used for this project. A “path matrix” recording the shortest number of steps between any two nodes i and j, can be generated and saved.

*path\_matrix = johnson.all.pairs.sp(inv\_graph)*

Fitting a histogram to the data indicates that relationships with a path less than two are likely to be meaningful. Paths of two or less comprise just over 10% of the total path lengths. Frequency increases dramatically in the distribution as we move above 2.0.



**Subgraph of an Adjacency Matrix**

Given that SVB has knowledge of high-value investors, it may be valuable to observe how “close” a given company is to these investors. A function has been designed that accepts the path matrix, a list of person nodes currently associated with a company, and investors of influence.

The function computes summary statistics on the relationships the current people have with influential investors. This includes computing the mean, interquartile range, and counting “significant” relationships.

Further work was performed using the degrees function to assess node-by-node centrality. Adjacency and densely connected subgraphs were also plotted. The adjacency plot above indicates how dense these networks can get. The most densely connected graphs were too complex for R to plot.

**B. Exponential Random Growth Model**

As previously described ERGM’s fit distributions to structural features in the graph, assuming homogeneity unless node attributes are set to indicate otherwise.

Relational data lists and adjacency matrices can both be used to create a stagnate network object, which supports the ERGM and BERGM packages.

ERGM sets up the MLE for the network and enacts a gradient descent. Full training on a model of this size will take considerable time. The gradient descent estimates the parameters of the underlying distributions.

The structural features of *edges*, *mutual*, and *triangle* were used to compare models, with the best model selected based on AIC.

*model\_from\_adj\_mut\_tri = ergm(net\_from\_adj~edges+mutual+triangle)*

Coefficients cannot currently be shown, as they may constitute private information. However, the coefficients express the log likelihood that, for each instance of the given feature, a node may create a new relationship with another random node.

For instance, for each edge that a node participates in, it’s likelihood of forming an edge with some other, random node, would increase by 1 times the exponentiated coefficient.

Edges, mutual ties, and triangle features were all found to be statistically significant, with p-values approaching 0.

A numeric approach may also be taken using the BERGM package. This sets up a modelin the same manner as ERGM. However, BERGM simulates the graph repeatedly and draws distributions from the simulations. Therefore it also accepts variables for iterations and number of chains.

The coefficients found via the numeric and analytical processes were found to be of practical similarity.

**V. Conclusions and Further Work**

The graph appears to be generally sparse, with some extremely dense subgraphs. Path distances of 2 appear significant, and various measurements of how close a given company is to influential investors can be evaluated.

ERGM’s can successfully fit significant structural features to an investor net.These values can be used as a posterior predictive, but that will only hold significance after normalization and particularly after additional data is added to the model.

Further work with ERGM’s will include the use of nodal covariates, distance, and weights. The values from the path matrix may serve as a proxy for distance and weight counts have been computed.

Nodal features, such as industry preferences or a tendency to lend in early or late stages, can also be added to differentiate between nodes. It is relatively unlikely that a founder at an incubator will receive seed funding from an established, late-stage venture firm.

With sufficient inclusion of node and edge data, a prolonged gradient descent should produce a model capable of predicting the future development of investor social networks. Combined with belief propagation and other decision-making heuristics, this will enable long-term planning to reflect the changing landscape of investor social networks.