A Graph Analysis of Investment Patterns

Samhita Karnati

Advisor: Prof. Andrea LaPaugh

Motivation

"Raising capital when nobody knows who you are is one of the most discouraging parts of starting a company."

- Garrett Lord, CEO/Co-Founder of Handshake

- Entrepreneurship supposedly democratic
- Funding often based on who a founder knows
  - Very difficult for first-time founders to efficiently seek out funding
- Discover patterns in investment to determine which investors are likely to fund a given venture

Goal

Investigate investment patterns in startups

- Fine patterns
  - Can we predict specific company-investor relationships based on past investments?
- Coarse patterns
  - Do investment patterns relate to company category?

Related Work

- Financial analysis for later-stage events[1] – general trends
  - Mergers & Acquisition prediction
  - IPO prediction
  - Dataset limited to firm's portfolio/dealflow
- Categorical and non-financial features for startup similarity and prediction
  - What features are common among successful startups
  - Do startups in similar categories (i.e. transportation, healthtech, edtech, etc.) have similar features?
  - Event prediction using categorical features[2] – specific events

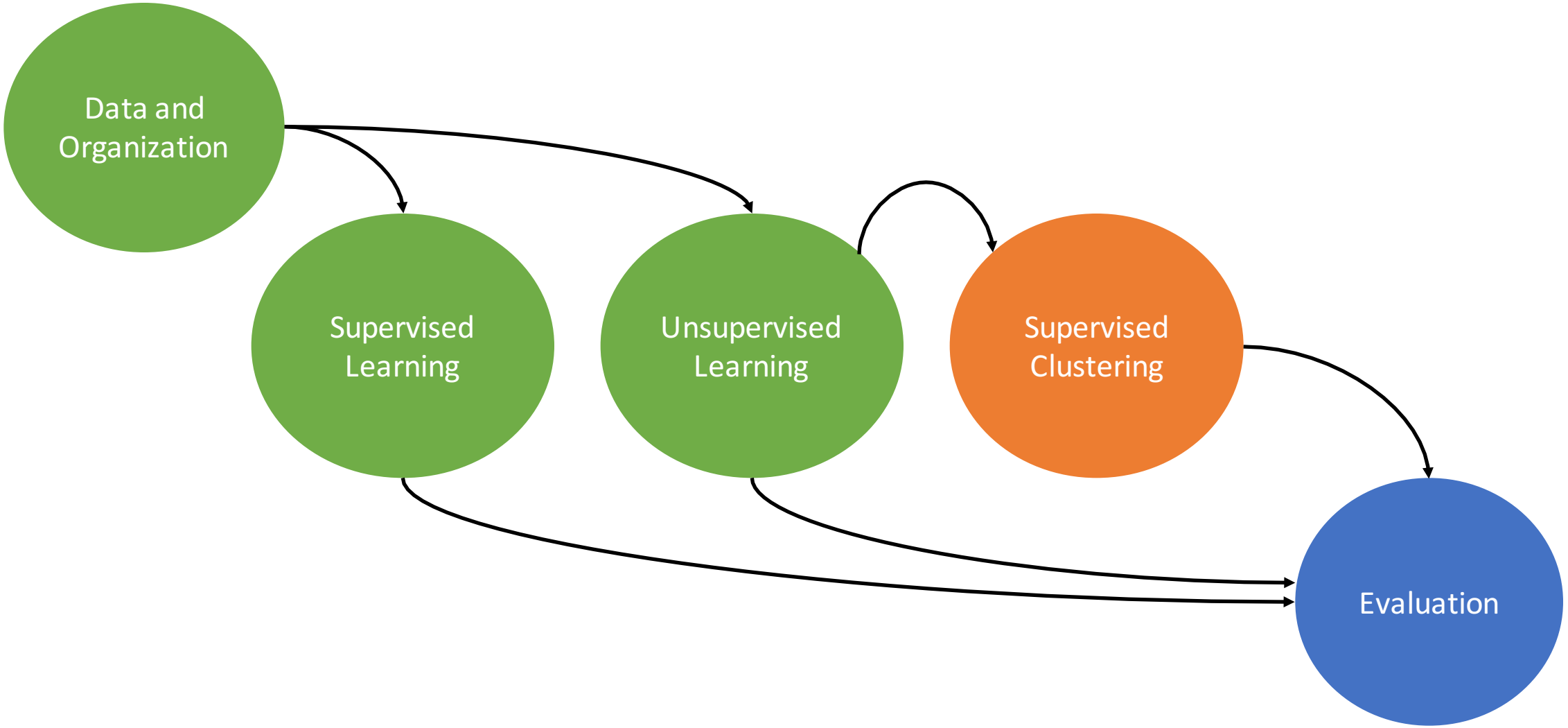
[1] Surr, Pierre-Alain. "2016 Technology Industry Trends." Strategy&. Strategy& and PwC, n.d. Web. 3 Oct. 2016.

[2] Xiang, Guang, et al. "A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch." ICWSM. 2012.

Approach

- Use objective, numerical information to
  - Predict specific events – funding
  - Investigate correlation with general patterns – company category
- Use more representative dataset – crunchbase
  - Do not constrain to particular firm or market

Implementation



Implementation: Data and Organization

- Data from crunchbase
  - 74,339 companies with complete funding and categorical data
- Use funding rounds csv, storing
  - Company name
  - Funding rounds
  - Investors
  - Geographic location
  - Category tags
- Create a digraph
  - Nodes: companies and investors
    - Attributes: Category and location
  - Edges: investments
    - Weights: Dollar amount transacted

Implementation: Supervised Learning

Link Prediction: company-investor relationships

- Split data: 90% training, 10% testing
- Link-prediction algorithms
  - Modified Jaccard similarity score

$$J(A,B) = \frac{(A \cap B)}{(A \cup B)}$$

where A = company's neighbors' neighbors
B = investor's neighbors

  - Adamic-Adar index
    - Similar to Jaccard, but weights rare features more heavily
  - Preferential attachment score
    - Multiply number of investors of company by number of companies the investor has invested in

Results & Evaluation: Supervised Learning

Full evaluation not done, but initial, cursory look over results makes sense

- Jaccard seems to work better than Preferential Attachment
- Jaccard seems to work better than Adamic-Adar

Plan going forward

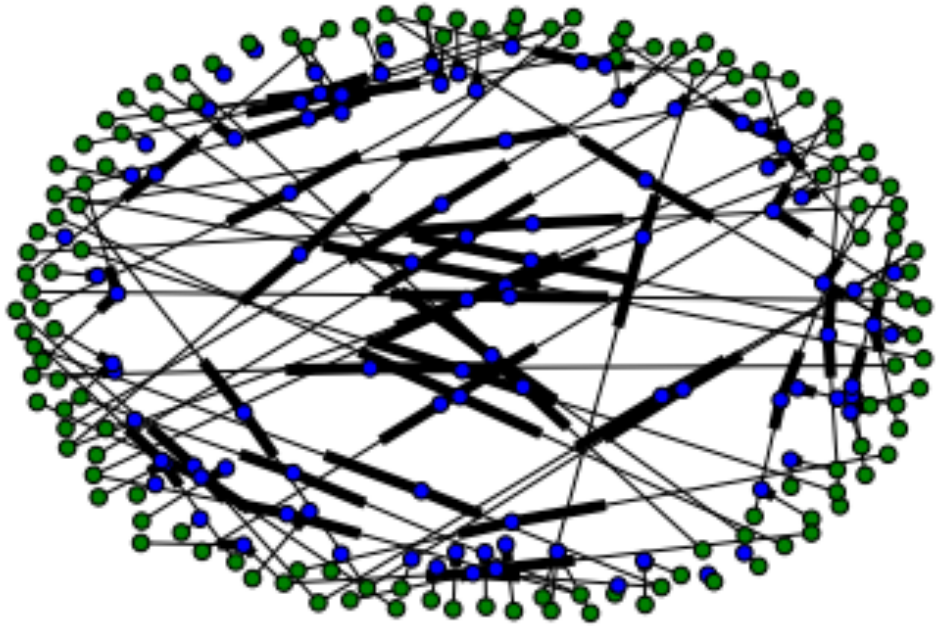
- ROC curves: receiver operating characteristic curves
  - Illustrates the performance of a binary classifier system
  - True positive rate against false positive rate at various threshold settings
  - Goal: maximize area under the curve

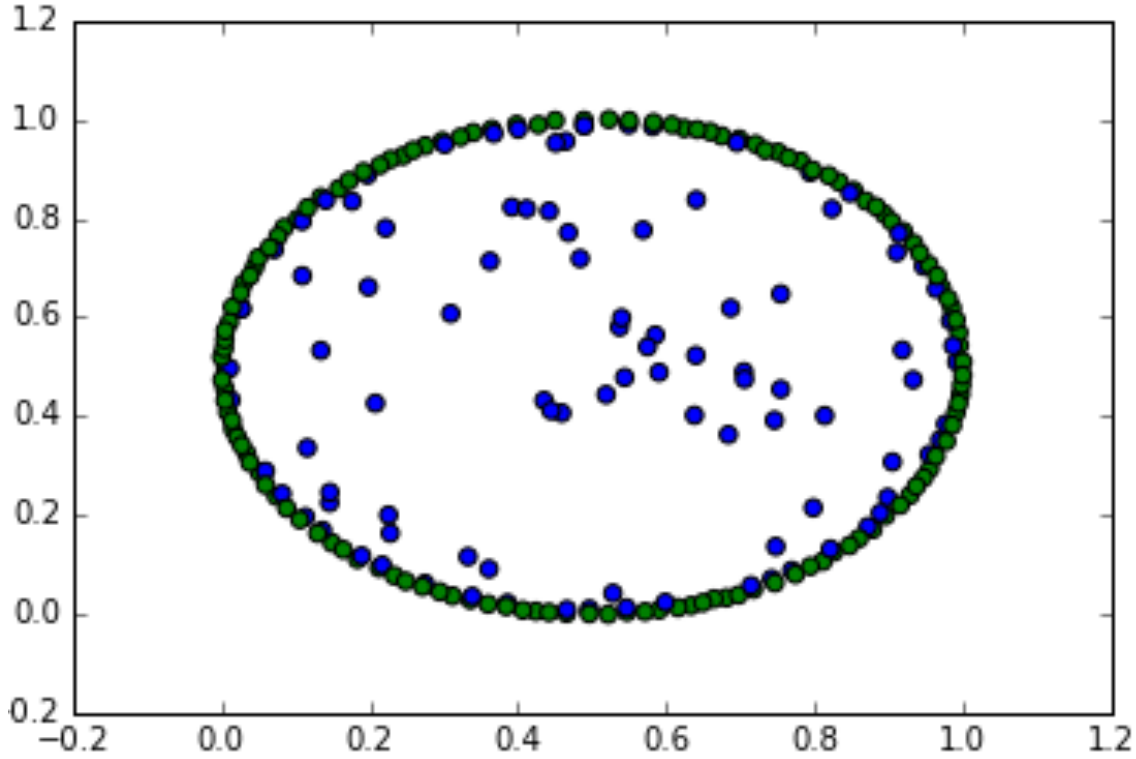
Implementation: Unsupervised Learning

Clustering and investigating category within clusters

- K-means clustering
  - Limit to SF-region companies
  - Hypothesis: companies in the same category have similar investing patterns and would therefore be in the same cluster
  - 550 unique first-order categories – use this for cluster number
- Cross validation
  - Fine-tuning cluster size parameter using k-fold cross validation

Results & Evaluation: Unsupervised Learning



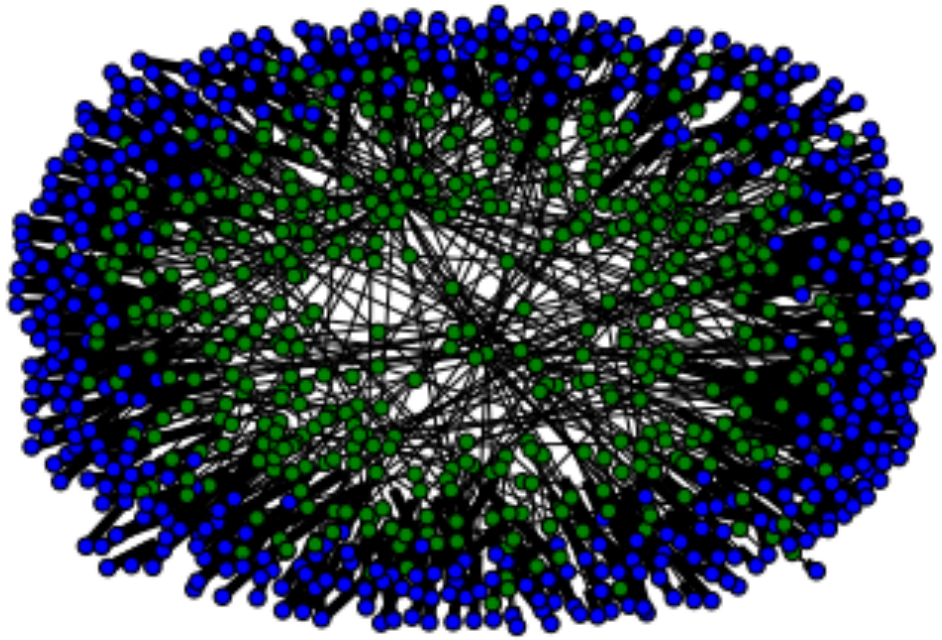


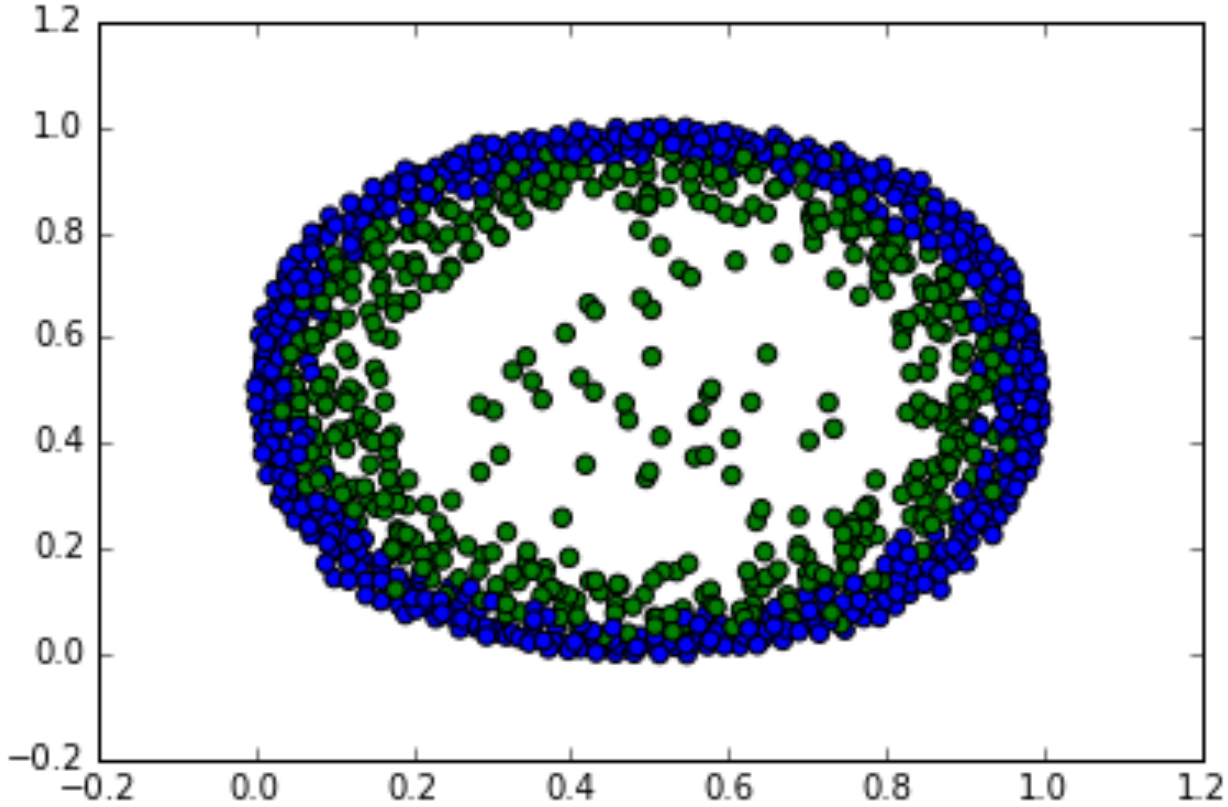
100 random companies with their investors

Blue nodes: companies

Green nodes: investors

Results & Evaluation: Unsupervised Learning





500 random companies with their investors

Blue nodes: companies

Green nodes: investors

Results & Evaluation: Unsupervised Learning

Plan going forward

- Evaluate k means with 550 clustering
  - Metric 1: are all the companies in the cluster of the same category?
  - Metric 2: are all the companies of a given category captured within a cluster?
  - Take out VCs from clusters
- K-fold cross-validation
  - Do VCs form their own clusters?
  - Is there an optimal number of clusters to capture industry-specific VC clusters?

Supervised Clustering

- Based on work by Finley and Joachims from Cornell
- Structural support vector machine (SSVM) algorithm for supervised k-means learning problem
  - Directly optimizes a similarity measure to maximize cluster accuracy
- Training set: form of sets of items with desired partitioning
- Two options
  - Spectral relaxation
  - Traditional k-means algorithm
- Do similar evaluation as with unsupervised learning

Future Work

- Get more than just binary information
  - Predict how much money will flow, not just if it will
- Add other attributes other than just geographical location and category
  - Investors typically fund the same founder in multiple ventures
  - Series information – size of company

Acknowledgements

- Prof. Andrea LaPaugh
- Peers in my IW seminar
- Crunchbase