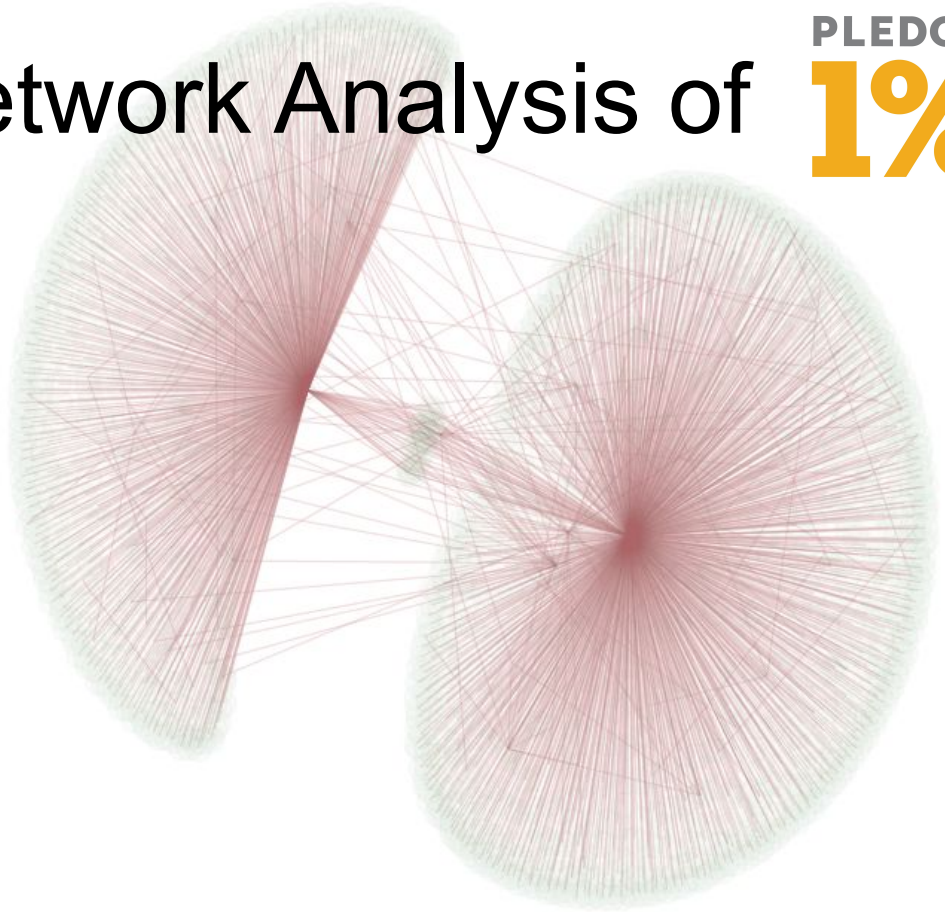
Network Analysis of PLEDGE 1% with Crunchbase



Aditya Mengani
Ognjen Sosa
Sanjay Elangovan
Song Park
Sophia Skowronski

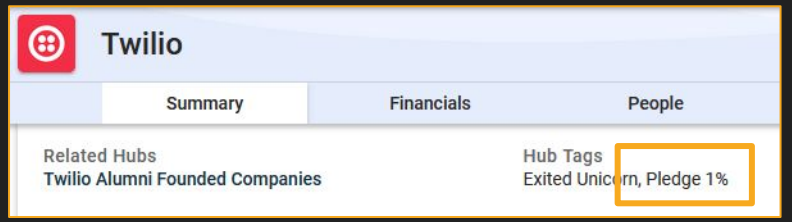
Pledge 1%

- Movement of corporate philanthropy via a membership model
- More than 12,000 companies in over 100 countries committed to giving back
- Membership requirements: CEO/Senior Exec sign-off & website
- Pledge types
  - 1% of equity
  - 1% of staff time
  - 1% of product
  - 1% of company profit

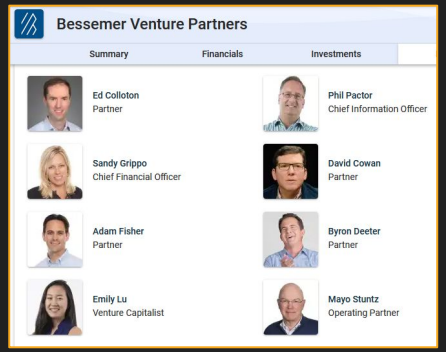
FAST COMPANY

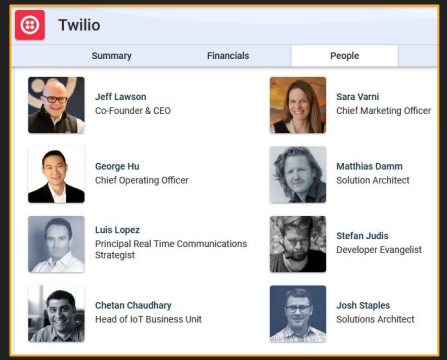
2017 MOST INNOVATIVE COMPANIES

Using Crunchbase



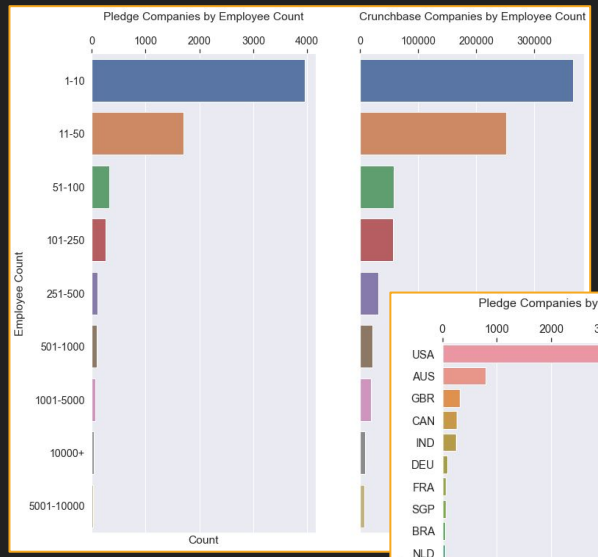
- Data sourced with Enterprise License Agreement
  - +1M organizations
  - Contains company information, leadership, investments and funding, etc.
  - Pledge 1% companies are tagged in Crunchbase



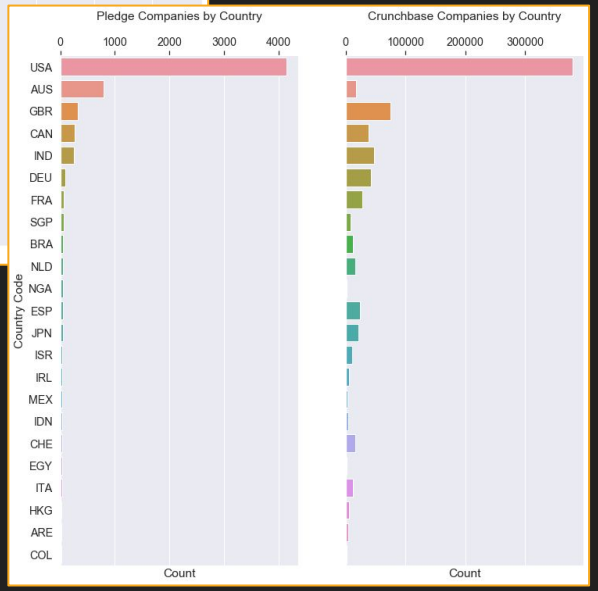


[Crunchbase]

Pledge 1% Companies Key Insights



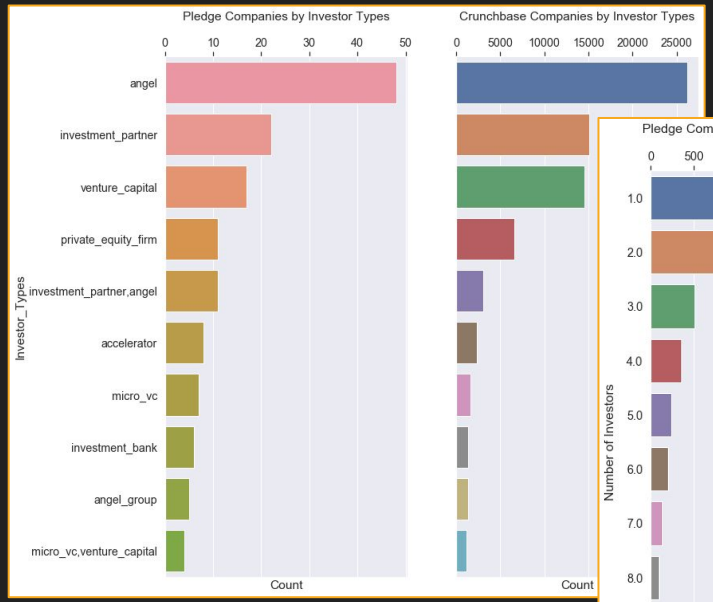
… are predominantly US based and ...



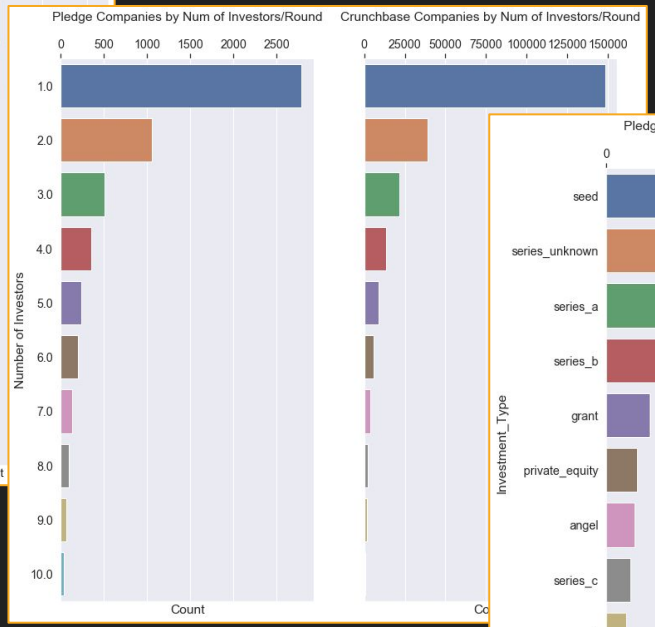
When compared to Crunchbase, Pledge companies tend to have smaller number of employees ...

… are more heavily comprised of tech co's.

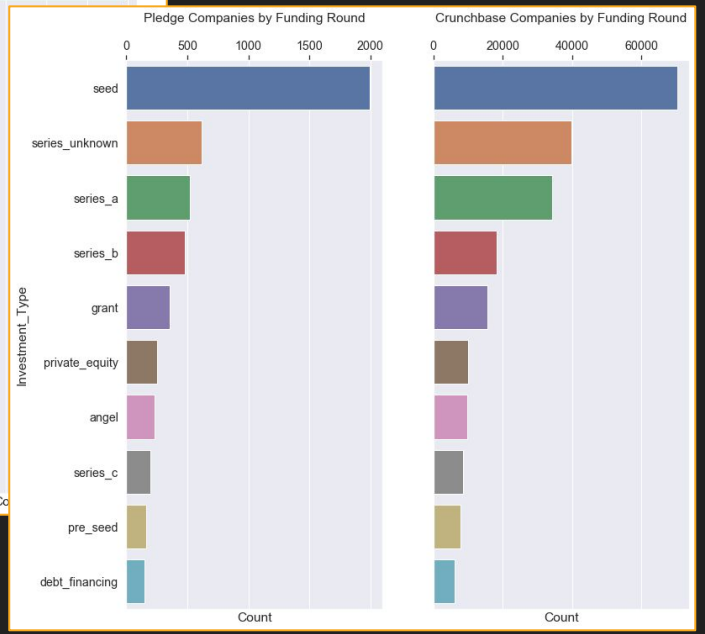
Pledge 1% Companies Key Insights



... that step in early in seed rounds ...



... with a larger # of investors participating.



When compared to Crunchbase, Pledge companies tend to source capital from angel investors ...

What we found

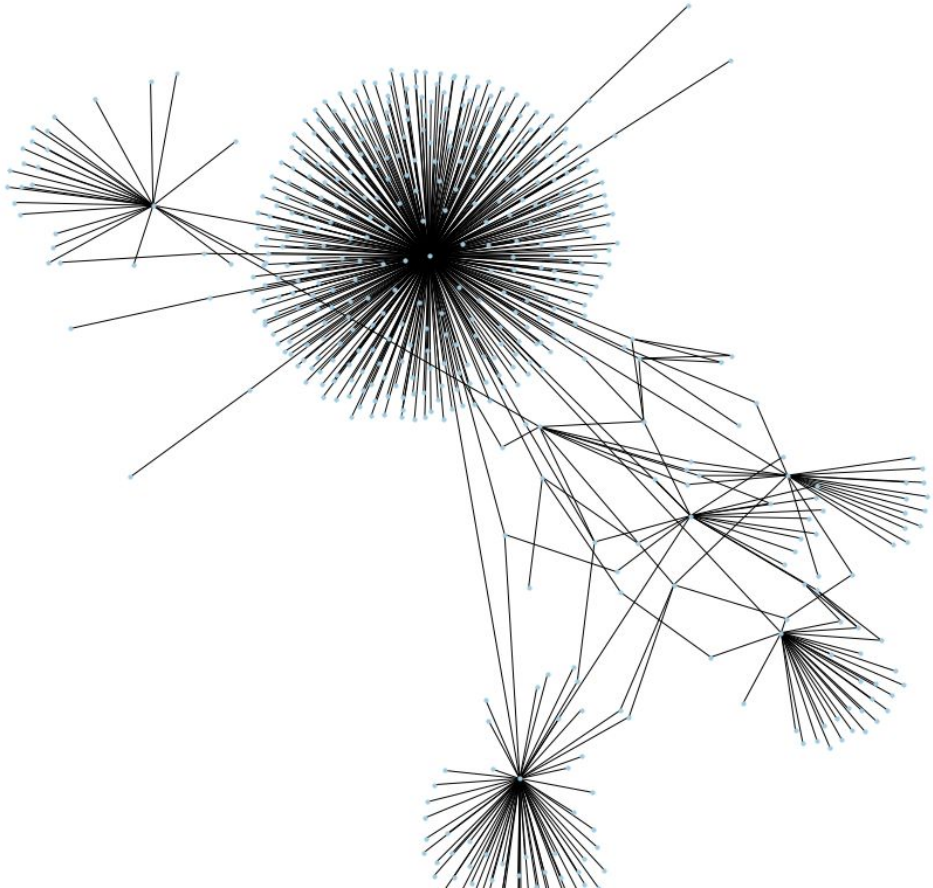
- **Pledge 1% has a unique footprint in Crunchbase**
  - Industry
  - Country
  - Age
  - Total funding
  - Rank
- **Using the variables above, we received modest accuracy scores in predicting whether or not a company was a Pledge 1% member**
  - **Baseline:** LRR 50% | BNB 68% | KNN 62%
- **Focus of this presentation**: leverage network information to improve baseline scores

Creating the graph

- Decisions
  - Representing the relationships and their significance
  - Reducing overall size of Crunchbase for sampling → 4 or 5 degrees away
  - Graph features on sampled network → 3 degrees away

Primary relationships

Direct jobs and investments

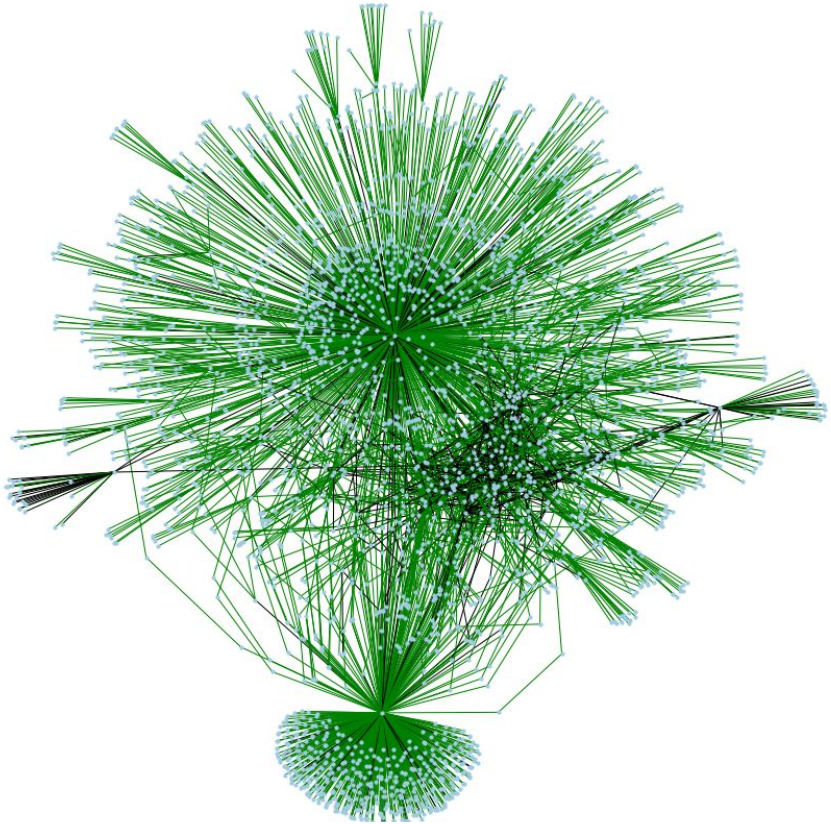


Google (via Crunchbase)

Secondary relationships

What's added in green

Previous employment & indirect investments

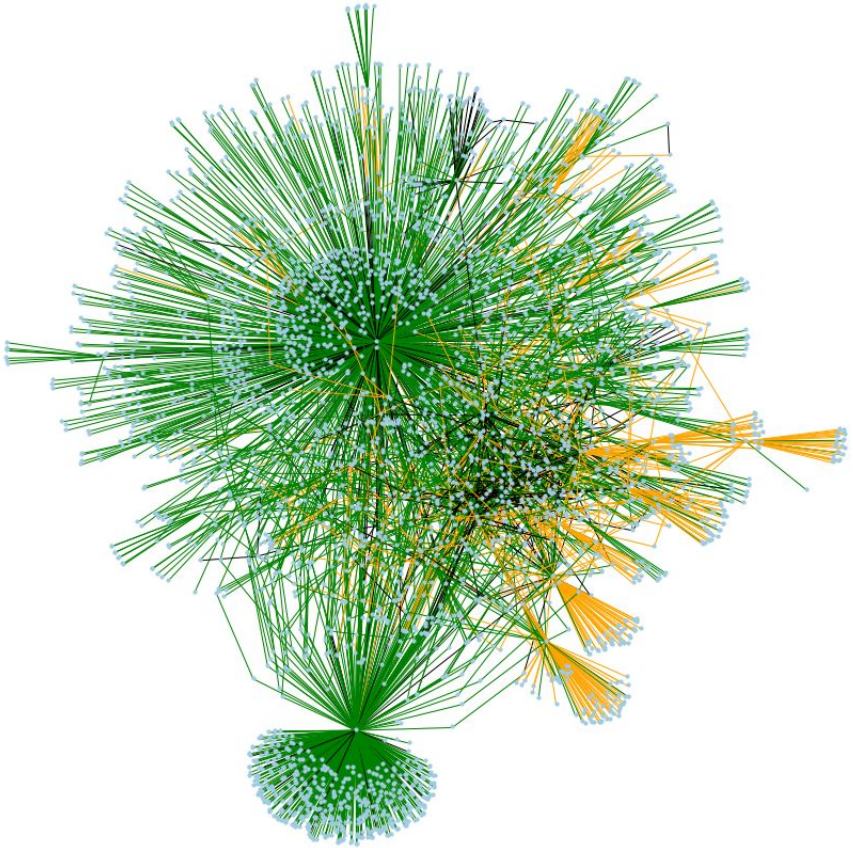


Google (via Crunchbase)

Tertiary relationships

What's added in orange

Coworkers and partners at the same firm



Google (via Crunchbase)

Overall graph features

- Size of networks
  - Crunchbase G(V: **1.3M**, E: **2.1M**)
  - Pledge 1% G(V: **61K**, E:**121K**)
  - Model G(V: **385K**, E: **733K**)
- Density of networks
  - Crunchbase 0.00000250475
  - Pledge 1% 0.0000649082
  - Model 0.00000988583

Graph pipeline processing for modeling

- 50 graphs x 2 neighborhood sizes = 100 total
  - 4 degrees away from Pledge 1% companies
  - 5 degrees away from Pledge 1% companies
- 100 x 15 minute processing step
  - Sample organizations from one of the two lists
  - Reduce Crunchbase graph to 3 degrees around sample
    - ~12K records each (half/half +/-)
  - Calculate graph features
    - Pagerank, weighted pagerank, shortest path, weighted shortest paths, in degree/out degrees, k-core decomposition

Our graph analytics struggle session

- turicreate saved the day
- Otherwise, needed GPU support
- Many NetworkX methods inaccessible due to timing constraints
- RAPIDS cuGraph: GPU-accelerated graph algorithms

graph_analytics

- connected components
- degree counting
- graph coloring
- k-core
- label propagation
- pagerank
- shortest path
- triangle counting

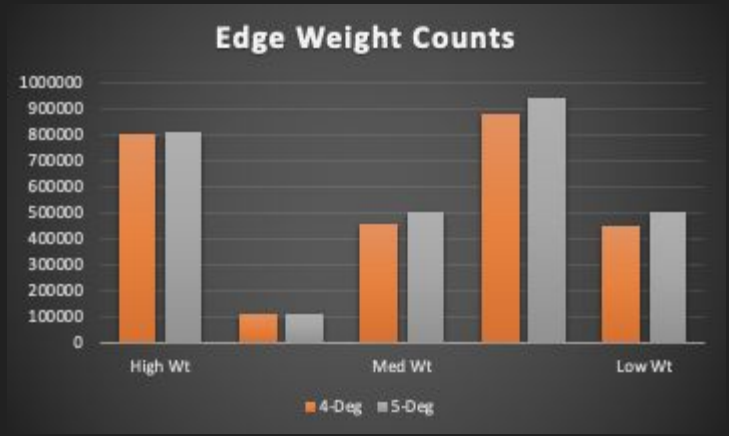
RAPIDS **cuGraph - GPU Graph Analytics**

build passing

The RAPIDS cuGraph library is a collection of GPU accelerated graph algorithms that process data found in GPU DataFrames. The vision of cuGraph is *to make graph analysis ubiquitous to the point that users just think in terms of analysis and not technologies or frameworks*. To realize that vision, cuGraph operates, at the Python layer, on GPU DataFrames, thereby allowing for seamless passing of data between ETL tasks in cuDF and machine learning tasks in cuML. Data scientists familiar with Python will quickly pick up how cuGraph integrates with the Pandas-like API of cuDF. Likewise, users familiar with NetworkX will quickly recognize the NetworkX-like API provided in cuGraph, with the goal to allow existing code to be ported with minimal effort into RAPIDS. For users familiar with C++/CUDA and graph structures, a C++ API is also provided. However, there is less type and structure checking at the C++ layer.

Applying weights

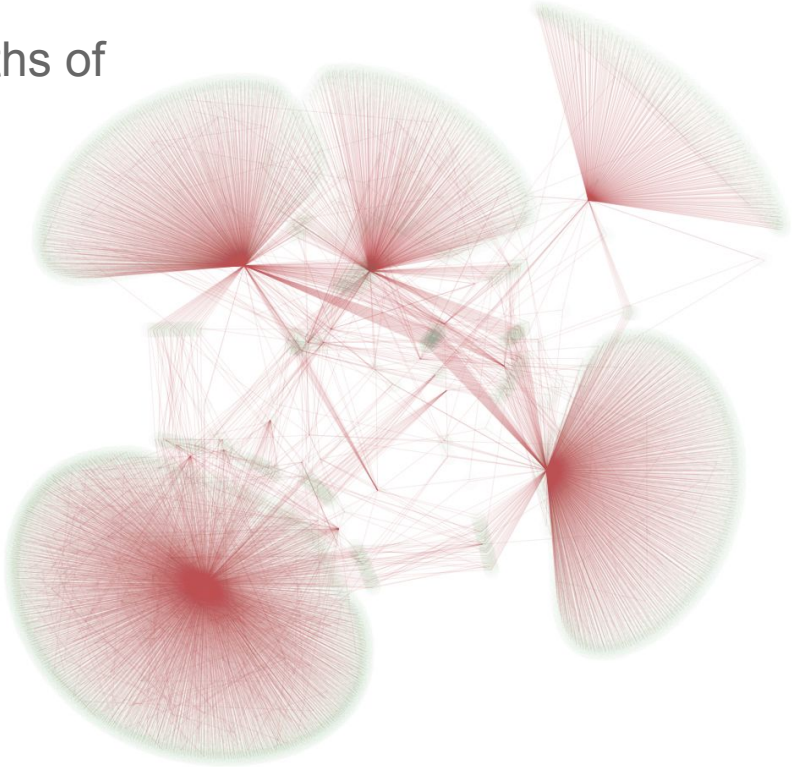
- Network relationships can be influenced by number of factors
- Graph edges can be weighted differently based on perceived ‘real-world’ connectivity
- Our methodology: from baseline EDA and intuition, edge weights assigned based on importance



Graph feature highlight: pagerank

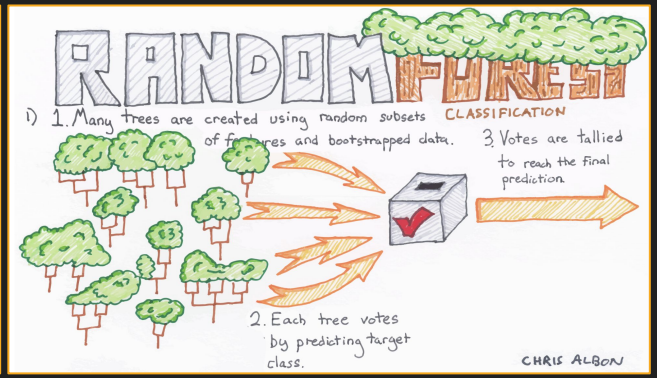
Weighted edges represent real-world strengths of relationships

Rank	Pagerank	Weighted Pagerank
1	Google	Salesforce
2	SAP	KPMG
3	Yahoo	Zuora
4	Salesforce	Techstars
5	KPMG	Box
6	Flexport	SAP
7	Samsung	Docusign
8	Techstar	Slack
9	Atllasian	Google
10	Twilio	Yahoo



Model pipeline

- The following results show model in different contexts and accuracy scores
  - For LRR, SVM, NB, Decision Trees, Random Forest, XGBOOST
  - With/without graph features
  - With/without other feature treatments
  - With/without expanded graph network



Model Results

Graph: k-core + min shortest path + shortest paths + degrees (in/out) + pagerank
Baseline: age + industry + employee count + country + rank + total funding
Baseline-R: age + industry + employee count + country

DEGREE	TYPE	LRR	KNN	BNB	DCT	XGB	RMF	SVM
4 deg from P1	Graph	0.684	0.675	0.667	0.709	0.735	0.734	0.673
	Baseline-R	0.710	0.702	0.612	0.762	0.767	0.800	0.751
	Baseline	0.710	0.713	0.596	0.775	0.877	0.820	0.724
	G+Baseline-R	**0.789**	**0.935**	**0.699**	**0.943**	**0.978**	**0.993**	**0.876**
	G+Baseline	**0.794**	**0.957**	**0.715**	**0.972**	**0.994**	**0.996**	**0.899**
5 deg from P1	Graph	0.690	0.688	0.675	0.714	0.726	0.734	0.687
	Baseline-R	0.718	0.697	0.641	0.739	0.787	0.771	0.736
	Baseline	0.710	0.688	0.629	0.790	0.855	0.836	0.757
	G+Baseline-R	**0.803**	**0.935**	**0.743**	**0.941**	**0.985**	**0.996**	**0.878**
	G+Baseline	**0.813**	**0.950**	**0.732**	**0.930**	**0.977**	**0.993**	**0.902**

Δ Model Results

Graph: k-core + min shortest path
Baseline: age + industry + employee count + country + rank + total funding
Baseline-R: age + industry

DEGREE	TYPE	ΔLRR	ΔKNN	ΔBNB	ΔDCT	ΔXGB	ΔRMF	ΔSVM
4 deg from P1	**Graph**	**0.114**	**0.295**	**0.081**	**0.125**	**0.139**	**0.148**	**0.087**
	Baseline-R	**0.05**	**0.051**	**-0.036**	**0.033**	**0.05**	**0.073**	**0.037**
	Baseline	-0.014	0.013	-0.049	-0.015	-0.008	-0.006	-0.006
	G+Baseline-R	0.062	-0.001	0.016	-0.013	-0.001	0.005	0.023
	G+Baseline	0.029	-0.003	0.032	-0.013	0.003	0	0.001
5 deg from P1	**Graph**	**0.108**	**0.115**	**0.09**	**0.11**	**0.156**	**0.149**	**0.102**
	Baseline-R	**0.044**	**0.014**	**-0.004**	**0.01**	**0.05**	**0.04**	**0.043**
	Baseline	0	0.023	0.015	-0.037	-0.011	0.005	0.02
	G+Baseline-R	0.038	-0.005	0.037	0.069	0	0.004	0.036
	G+Baseline	0.015	0.003	0.057	0.027	0.004	0	0.002

What's next?

V2 Time series

- Incorporate time as a feature
- Link prediction: predict the changes of edges or nodes of networks over time

V3 Knowledge graph

- Derive semantic information from relationships to enhance link prediction modeling