Project Proposal

Group 131

CSE6242—Fall 2022

What are we trying to do?

Use Newman's Ground Truth Algorithm[1] (NGTA) on a patent network to filter prolific inventors from network noise.

Because the patent network is noisy, we are skeptical that it accurately reflects reality. NGTA will allow us to quantify our confidence that each edge is "real" and not noise. NGTA will produce statistics that quantify the reliability of the network.

[1] Newman, Mark EJ. "Network structure from rich but noisy data." *Nature Physics* 14.6 (2018): 542-545.

How is it done today, and what are the limits of current practice?

- Noisy graphs are often filtered on weights. This **assumes a linear relationship** between weights and confidence[1].
- Robust ML approaches use semi-supervised learning and inference models[2], which are **computationally expensive**.
- Ensembles methods (generative stochastic graph models[3]) **rarely have closed-form solutions**, requiring simulations.

[1] Namata, Galileo Mark, Stanley Kok, and Lise Getoor. "Collective graph identification." Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011.
[2] Casiraghi, Giona, et al. "From relational data to graphs: Inferring significant links using generalized hypergeometric ensembles." International conference on social informatics. Springer, Cham, 2017.
[3] Faysal, Md Abdul Motaleb, and Shaikh Arifuzzaman. "A comparative analysis of large-scale network visualization tools." 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018.

What's new in our approach?

- NGTA has **never been executed on patent network data**.
- NGTA will enable us to quantify confidence in each edge’s existence in the network, then filter the network dynamically by confidence levels.
- NGTA **extracts inventors who are “real” contributors**, and removes those less important to a company's patent portfolio.

Who cares?

In an ever-competitive war for talent, companies will increasingly turn to data-driven recruiting strategies.

Patent networks are directly related statistically to a firm's acquisition and dispersion of new technology[1]. Identifying true innovators allows firms to poach those people for competitive advantage. NGTA allows recruiters to confidently identify the best individuals to pursue.

[1] Duguet, Emmanuel, and Megan MacGarvie. "How well do patent citations measure flows of technology? Evidence from French innovation surveys." Economics of innovation and new technology 14.5 (2005): 375-393.

What is the impact, how can we measure it?

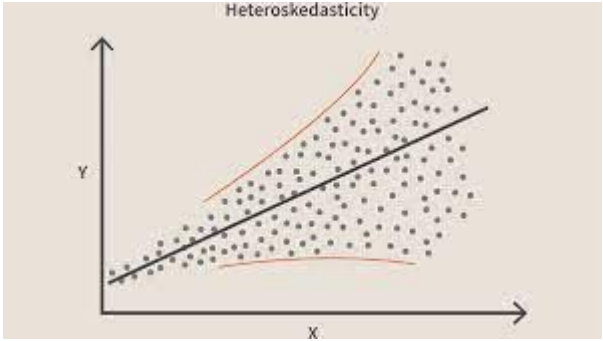
The Impact:

- Enables organizations to increase their confidence in who to target for potential hires based on their patent contributions.

Metrics:

- NGTA converges
- Network statistics between training and holdout set are similar

What are the risks and payoffs?



Risks:

- NGTA doesn’t converge to stable values, which we’ll resolve by segmenting the network.

Payoffs:

- Increase business intelligence by extracting performant individuals that recruiters can target.

How much will it cost?

$0.

The dataset and algorithms will all fit within the free tier of Google Cloud Processing (GCP).

How long will it take? Who will do what?

1. Development (4 weeks)
    a. Tool Setup (Everyone)
    b. Data Acquisition (Sang Yoon and Erik S)
    c. Algorithm Implementation (Tylor)
    d. Front End Development (Sandro & Rick)
2. Testing (2 week)
    a. Front End Testing (Everyone)
3. Progress Report (2 week)
    a. Poster (Everyone)
    b. Individual Video Presentations (Everyone)
    c. Final report write-up (Everyone)

How will progress be measured?

- **Midterm:** NGTA implemented with limited visualization capabilities on a subset of the patent network
- **Final:** NGTA implemented with full interactive visualization capabilities on the entire patent network