Patently True

Introduction

Patent data is inherently noisy, with many inventors authoring many patents. This makes it difficult to determine those who actually contribute meaningful work. Furthermore, the data is quite large, making manual filters difficult - if not impossible - to implement efficiently or effectively. We propose using Newman's Ground Truth Algorithm (NGTA) to efficiently filter the noisy patent data and identify prolific inventors. NGTA's outputs can then be used for recruiting purposes, to hire top talent.

Problem Definition

Using NGTA on a patent network extracted from the PatentsView database, we will 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.

Survey

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

Noisy graphs are often filtered on weights. This assumes a linear relationship between weights and confidence. Robust ML solutions, such as Namata's Coupled Collective Classifiers (C3)[2], can perform network inference (extract the "real" graph from the observed). These approaches use semi-supervised learning and inference models[2], which are computationally expensive. While some semi-supervised models process noisy graphs, they are usually limited to clustering[12]. Others use ensembles (generative stochastic graph models)[3]. These are rarely closed-form, requiring simulations. A common goal with noisy networks is community detection,[15, 16] which does not assess individual components. A method developed by Dr. Du (Beijing U of T) processes patent networks to find influential inventors[18]. His is more efficient than prior approaches, but does not distinguish ground truth. If non-noisy patent data is acquired at scale, comparisons between Du's and Newman's methods would be insightful. In 1999, US patents contained 3.8 million nodes. This data is rich, with attributes for inventors, assignee, etc.[5] This richness results in noise, with innovators hidden by peripheral individuals. Older patent records can suffer from duplicate inventors, which present as false discoveries[6]. This data has historically been used to describe and forecast economic conditions, with common algorithms including shortest paths, and maximum flow.[5] Modern patent analysis has matured in several areas, including patent infringement, hotspots, and competitors[17]. Traditionally the talent industry relied on passive acquisition[8]. Today, "head-hunting" is limited to small groups. Patent data is rarely used, and only as a reference. NGTA brings patents to the forefront, enabling a data-driven recruiting strategy.

What's new in your approach? Why will it be successful?

In 2000, research on patent networks was primarily trend analysis of global structures, which fueled macroeconomic decisions[13]. Recently, patent analysis has been utilized to find impactful inventions and highly innovative companies[14] - largely fueling stock speculations. Never before has NGTA been executed on a patent network. NGTA will enable us to quantify confidence in each edge's existence in the network (where nodes represent inventors/ businesses, edges represent patents), then filter the network dynamically by confidence levels. After filtering, remaining connected nodes constitute significant inventors. Thus, NGTA extracts inventors who are “real” contributors, and filters out those who are less important to a company's patent portfolio. NGTA produces three network-level descriptive statistics. The true positive rate (low TP implies missing edges[1]), the false positive rate (low FP implies an edge is rarely observed where none exists[1]), and false discovery rate (low FDR implies observed edges are part of the network[1]). This implementation has a strong chance of success due to the algorithm being lightweight, with rapid convergence. NGTA can be run on tabular data, and does not require a graph database. The network will be instantiated using NetworkX, and visualized with Graphviz[4]. These open source packages offer robust, effective, and lightweight processing and interactive visualization capabilities.

Who cares?

Most empirical studies of networks take a naive view of structural data, where one assumes that the data are the network[1]. Accurate analysis and understanding of networked systems requires a way of estimating the true structure of networks from such rich but noisy data[1]. NGTA quantifies this discrepancy, and level-sets confidence in downstream outputs. Human capital has a profound impact on a firm’s strategy, outcomes, and performance. Furthermore, it is the most challenging resource for competitors to replicate.[9] 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[7]. Identifying true innovators allows firms to poach those people for competitive advantage. NGTA allows recruiters to confidently identify the best individuals to pursue.

If you're successful, what difference and impact will it make, and how do you measure them?

Successful implementation of NGTA will enable organizations to greatly increase their confidence in any business intelligence based upon the network. Recruiters can focus their efforts where the effort is worthwhile - the pursuit of valuable candidates. NGTA can be tested using a test set. If the resulting statistics are similar between the two, we can be confident that the algorithm reached true convergence.

What are the risks and payoffs?

This is a low-risk, high-reward implementation. One risk is that the network exhibits heteroskedasticity across industries, prohibiting NGTA from converging. If so, the network will be segregated by industry, with NGTA run upon sub-networks. Successful NGTA implementation upon patents results in business intelligence for the talent industry - extracting performant individuals, upon whom recruiters can focus.

How much will it cost?

Nothing. We can operate on the patent data in GCP within the limits of the free tier. While weekly updates to the US patent database can be hundreds of megabytes in size, the *patentpy* and *patentr* libraries allow for tidy and small records[10]. This greatly reduces the required data volume. Established conversion models between relational data and target graph data will further reduce the data volume[11].

What are the midterm and final "exams" to check for success? How will progress be measured?

The midterm consists of a small scale experimental run of NGTA on a subset of the network, with limited visualization. The final consists of a full run of NGTA on the full network, with interactive visualization.

Proposed Method

NGTA is the result of academic research by Dr. Mark Newman, published as *Network Structure from Rich but Noisy Data* in 2018. Newman discusses how data representing a network contains a hidden, "true" network structure that is obscured by noisiness. When a noisy network is not scrutinized, the true network structure remains hidden. NGTA is a lightweight algorithm that enables filtering of the network to edges that are statistically significantly likely to be a part of the true network structure. This likelihood is computed for and attributed to each edge in the network.

$$Q_{ij} = \frac{\rho\alpha^{E_{ij}}(1-\alpha)^{N-E_{ij}}}{\rho\alpha^{E_{ij}}(1-\alpha)^{N-E_{ij}} + (1-\rho)\beta^{E_{ij}}(1-\beta)^{N-E_{ij}}}$$

$$\alpha = \frac{\sum_{i<j} E_{ij}Q_{ij}}{N\sum_{i<j} Q_{ij}}$$

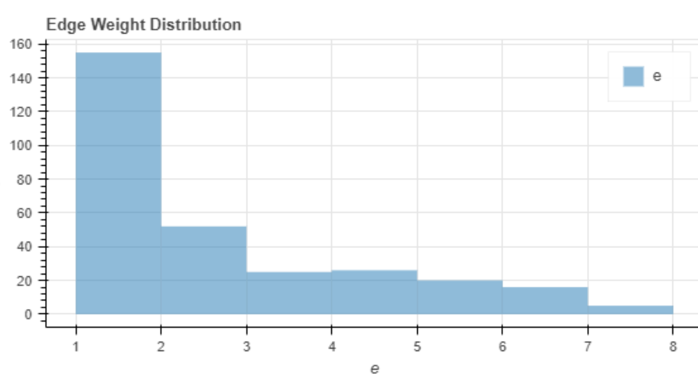
$$\beta = \frac{\sum_{i<j} E_{ij}(1-Q_{ij})}{N\sum_{i<j}(1-Q_{ij})}$$

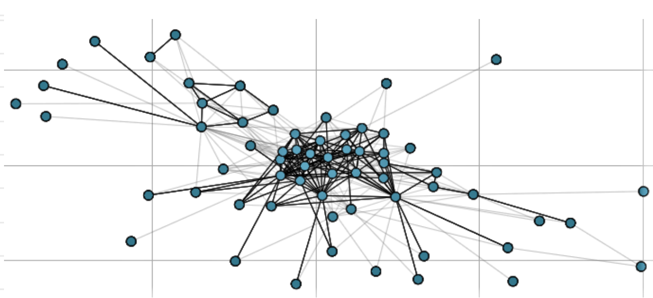
$$\rho = \frac{1}{\binom{n}{2}}\sum_{i<j} Q_{ij}$$

Each edge's likelihood (Q_{ij}) is based upon three network-level statistics: Alpha (True Positive Rate), Beta (False Positive Rate), and Rho (Prior Edge Probability). Each Q_{ij} value also depends on that edge's weight (e) and an overall measurement count (m). Alpha, Beta, and Rho are instantiated randomly. Then, an iteration begins. Within each iteration, each row's Q_{ij}, Alpha, Beta, and Rho are calculated. Iterations continue until Alpha, Beta, and Rho converge to stable values.

Alpha, Beta, and Rho *usually* converge rapidly to stable values. When convergence is not stable, this suggests one of two things: that the network exhibits significant heteroskedasticity across its surface, or that the distribution of the edge weights contains hills and valleys that create local minima and maxima. Fortunately, both of these issues are addressable. Heteroskedasticity is addressed by dividing the network into cohorts, and running NGTA upon each cohort. The patent network contains attribution for each patent, including industry. These attributes will serve as the means to subdivide the network.

If the distribution of edge weights presents hills and valleys, a transformation can be imposed upon that distribution, resulting in smoothing. Should this situation present itself, the specific and most suitable transformation will have to be discovered through trials.





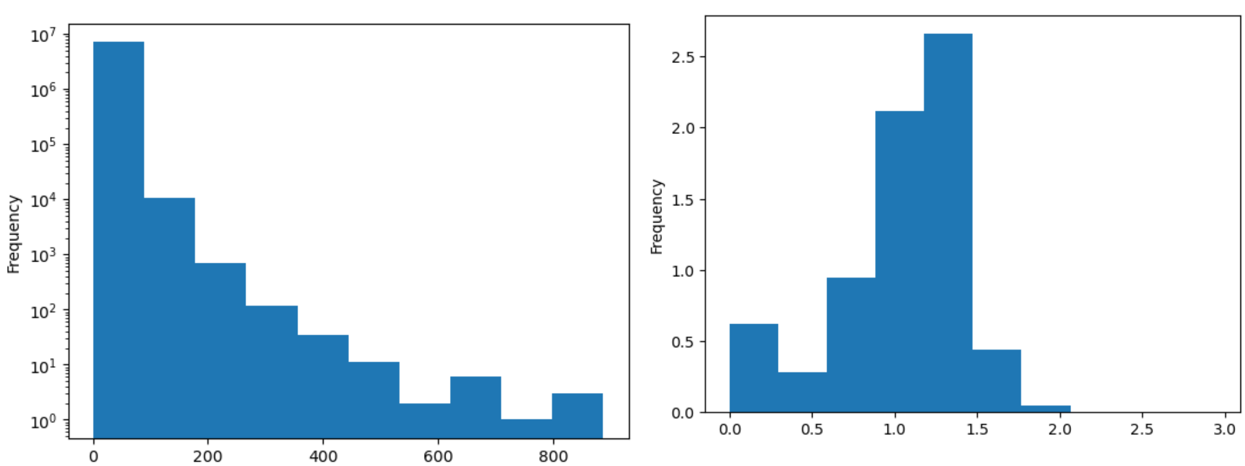
Newman's final network. Edge darkness corresponds to Q_{ij} values. Observe a core of nodes that are very likely connected to each other, and several peripheral nodes far less likely to be a part of the true network. Q_{ij} values range from above 0.999 to below 0.1.

NGTA is innovative in many ways. First, it is lightweight and unsupervised. It can operate on a large network (e.g. the patent network) without incurring large computational expenses. Secondly, the patent network is constantly changing, almost on a daily basis. Organizations tend to tackle noisy networks occasionally, due to the cost of conventional algorithms. NGTA can be re-run cheaply. Lastly, NGTA outputs network-level true positive and false positive rates. This makes NGTA predictive and diagnostic.

Experiments/Evaluation

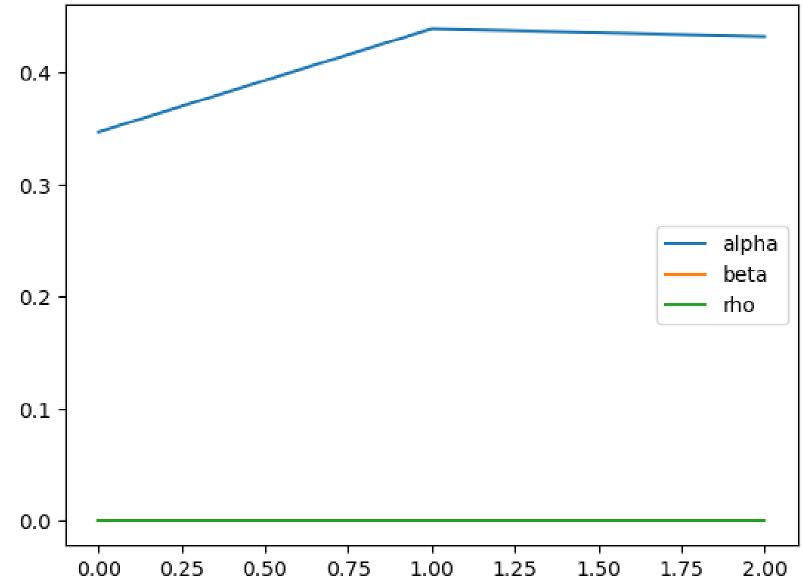
NGTA is a non-parametric, unsupervised algorithm; while it does not fit a distribution to data, *Patently True's* testbed is composed of four statistical tests, three of which are distribution-focused. First, validating smoothness of the frequency distribution of input data's edge values, which is a pre-requisite for the algorithm. Second, confirmation of convergence of each network-level meta-statistic across iterations. Failed convergences suggest violation of NGTA's assumption of low heteroskedasticity. Thirdly, bimodality of frequency distribution for Q_{ij} confidence metrics, confirming the strength of NGTA's classifications of edges into truth or noise buckets. And lastly, confirming non-linearity of the relationship between increased patent citations and Q_{ij} values - a linear relationship here suggests that NGTA found confidence to increase with citations on a one-for-one basis (in other words, provided no statistical insight to the noisiness of the patent network).

First test: citation distribution smoothness. *Patently True* utilizes patent citation counts as edge weights for NGTA. Patents are dominated by low citation-count nodes, which overwhelm the algorithm's ability to converge. A logarithmic scale showcases the distribution's extreme kurtosis.



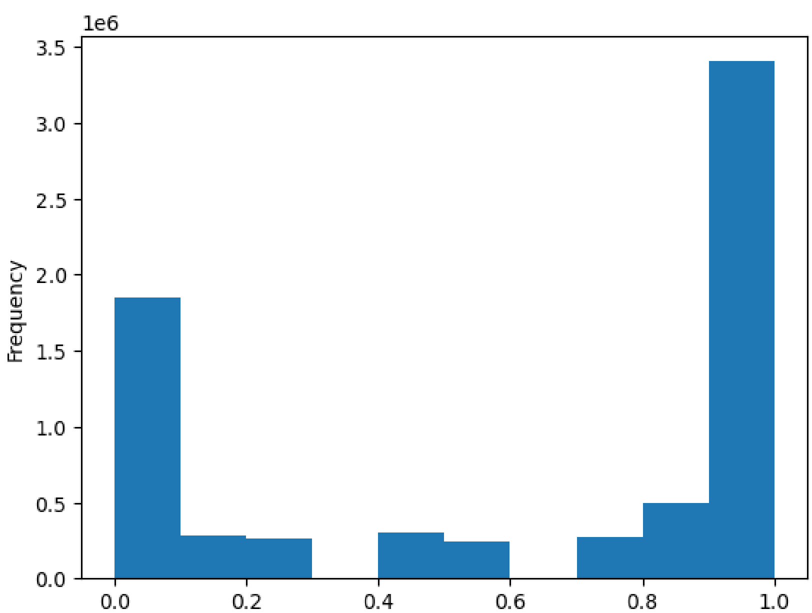
Left: original edge weight frequency distribution, viewed with logarithmic frequencies.
Right: edge weight frequency distribution after $\log_{10}$ transformation. Fortunately, this results in a distribution that NGTA is able to successfully process.

Test 2: Alpha, Beta, and Rho's convergences. NGTA's three network level meta-statistics (true positive = alpha, false positive = beta, and prior edge probability = rho) converge across iterations, which is verifiable by examining their iterative values.



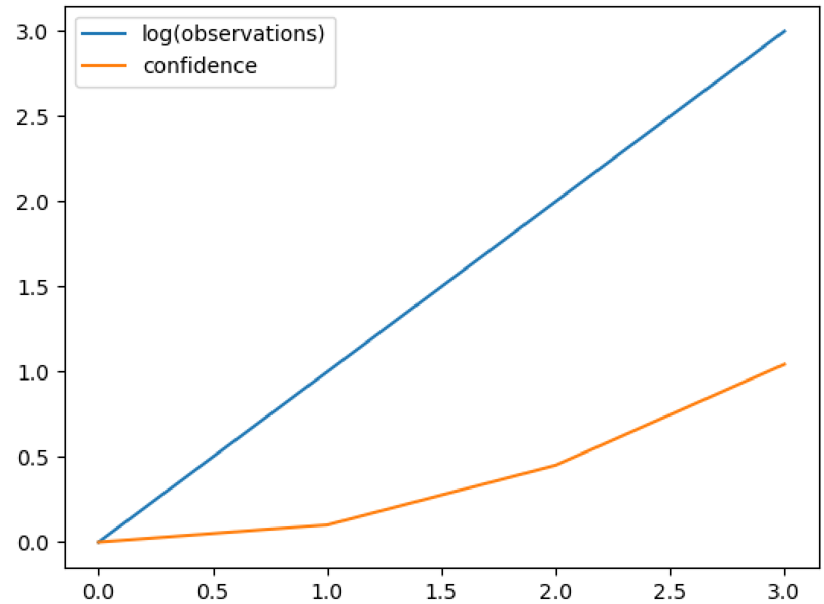
Shown here are successive alpha, beta, and rho values across iterations. Significant variance across iterations would convey NGTA's inability to converge. We see here that the algorithm converged rapidly for *Patently True*, much like for Dr. Newman's pilot experiment[1]. Rho and beta are both vanishingly small, and obscure each other in this plot.

For each edge in the network, NGTA computes a confidence metric (Q_{ij}), which quantifies confidence in whether that edge is part of the hidden, true network, or whether that edge is just noise. A Q_{ij} value at or near zero is almost certainly noise, and a Q_{ij} value at or near one is almost certainly real. Plotting the frequency distribution of these confidence metrics constitutes a statistical performance test for *Patently True*. Should we see a uniform distribution, or a distribution dominated entirely by a narrow range of values, or an abundance of values at or near 0.5, we would know that NGTA failed to produce meaningful confidence metrics for the patent network.



Frequency distribution for Q_{ij} confidence metrics - NGTA's edge-level outputs for *Patently True*. This bimodal distribution tells us that NGTA, overwhelmingly often, is very confident that a particular edge is either noise, or is part of the hidden true network. Rarely was the algorithm 'unsure' of the nature of an edge. These rare events are represented by the minority of Q_{ij} values at or near 0.5.

Final test: confirming non-linearity of the relationship between increased patent citations and Q_{ij} values. For NGTA's results in *Patently True* to be statistically insightful, a one-to-one relationship between observations and confidence cannot persist.



Shown here are both the NGTA-derived relationship between transformed (log) citation count and confidence in an edge (orange), as well as a 1:1 for reference (blue). We observe that NGTA finds the greatest increase in confidence between log(2) citations and log(3) citations, with much less between 0, 1, and 2. This validates NGTA's statistical insight within *Patently True.*

Conclusions and Discussion

Patently True reveals many surprising statistical insights into the patent network. The calculated true-positive rate of 43% implies that more than half of all true innovations are not even present in the network. The calculated false-positive rate is vanishingly small, which implies that the patent network is very well curated, with rare to non-existent events of mis-allocated patent authorship. Our last network-level meta-statistic, the false discovery rate, is 23% - meaning, NGTA has determined that nearly a quarter of all recorded innovation is noise.

Much like logistic regression, NGTA can be utilized as a classification algorithm, if a cutoff level is chosen. Using 0.5 as this cutoff, 38% of all patents in the network are not reliably indicative of innovation, while the remaining 62% do represent true innovation.

Patently True's implications for the talent industry almost cannot be understated. Knowing now that one quarter of the network is noise, that over one third of all patents do not indicate innovation, and that the network represents less than half of all cutting-edge work… these are humbling insights. With each and every edge in this network attributed with a statistically-derived confidence metric, talent acquisition will be able to step away from their previous, naive approach to the patent network. NGTA, and *Patently True*, teach us that while ground truth may be measurable, it is sometimes inferior to the abstract.

(All team members contributed a similar volume of effort to *Patently True*).

References:

1. Newman, Mark EJ. "Network structure from rich but noisy data." *Nature Physics* 14.6 (2018): 542-545.
2. 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.
3. Casiraghi, Giona, et al. "From relational data to graphs: Inferring significant links using generalized hypergeometric ensembles." *International conference on social informatics*. Springer, Cham, 2017.
4. 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.
5. Batagelj, Vladimir, et al. "Analyzing the structure of US patents network." *Data science and classification*. Springer, Berlin, Heidelberg, 2006. 141-148.
6. Li, Guan-Cheng, et al. "Disambiguation and co-authorship n etworks of the US patent inventor database (1975–2010)." *Research Policy* 43.6 (2014): 941-955.
7. 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.
8. Rose, Jacqueline A. "Building An Internet Recruiting Strategy For A Big Five Professional Services Firm." (1999).
9. Valenti, Alix, and Stephen V. Horner. "Leveraging board talent for innovation strategy." *Journal of Business Strategy* (2019).
10. Yu, James, et al. "Accessing United States Bulk Patent Data with patentpy and patentr." *arXiv preprint arXiv:2107.08481* (2021).
11. De Virgilio, Roberto, Antonio Maccioni, and Riccardo Torlone. "Converting relational to graph databases." *First International Workshop on Graph Data Management Experiences and Systems*. 2013.
12. Chang, Jui-Hung, and Hsiu-Chen Weng. "Fully used reliable data and attention consistency for semi-supervised learning." *Knowledge-Based Systems* 249 (2022): 108837.
13. Wu, Chao-Chan, and Ching-Bang Yao. "Constructing an intelligent patent network analysis method." *Data Science Journal* (2012): 011-003.
14. Chakraborty, Manajit, Maksym Byshkin, and Fabio Crestani. "Patent citation network analysis: A perspective from descriptive statistics and ERGMs." *Plos one* 15.12 (2020): e0241797.
15. He, Kun, et al. "Hidden community detection in social networks." *Information Sciences* 425 (2018): 92-106.
16. Newman, Mark EJ, and Michelle Girvan. "Finding and evaluating community structure in networks." *Physical review E* 69.2 (2004): 026113.
17. Abbas, Assad, Limin Zhang, and Samee U. Khan. "A literature review on the state-of-the-art in patent analysis." *World Patent Information* 37 (2014): 3-13.
18. Du, Yong-ping, Chang-qing Yao, and Nan Li. "Using heterogeneous patent network features to rank and discover influential inventors." *Frontiers of Information Technology & Electronic Engineering* 16.7 (2015): 568-578.