By submitting an application, I agree to the EGR submission requirements

First Names (Please list all co-authors)	William, Carma, Lizi
Surname/Last Name (Please list all co-authors)	Donovan, Khatib, Feng
Select Focus Area (Analytics, Economics, Or Relevance)	Analytics
Proposed Title	“Utilizing Taxonomy, Social Graph, and Keyword, Analysis, with Machine Learning Techniques, to construct a Predictive Data Model for understanding Endorsement Properties, and any correlation of Endorsement, and Endorser, to actual career success.”

A problem statement illustrating your research project

LinkedIn’s current “Skills and Endorsements” System (not to be confused with its “Recommendations” system) is broken – there is no clear correlation to, or predictability of, professional success and/or salary level, on the basis of A) having specific endorsements (KEYWORD ANALYSIS) or B) having endorsements from specific people who themselves are endorsed or have specific career or professional characteristics. (SOCIAL GRAPH ANALYSIS) It is therefore impossible to tell what are High Value Skills, and who has High Value Skills with High Credibility.

Furthermore, this system produces a fractured data set that requires normalization – for example, the same individual can be endorsed for “Consulting,” “Management Consulting,” “Project Management,” and “Strategy.” (T-1: TAXONOMY OF ENDORSEMENTS) Career Titles must also be normalized into Talent Models to fully understand Credibility on Endorsement. (T-2: TAXONOMY OF TITLES)

This research is sponsored by The Tesla List, and will aid in our understanding of High Value Skills and developing a Curriculum for developing those Skills. See Additional Details and Short Bio sections for more information.

A brief description of your solution to the research problem

Taxonomy Analysis:

T-1: Taxonomy of Endorsements: Using statistical classification to recognize patterns of endorsement.

Currently, the variables (Endorsements) in Linkedin are not good at determining class labels for unseen instances, reducing their overall value, and the ability to create taxonomies that are collectively exhaustive by level, and mutually exclusive by level. For example, one may see two closely related instances of Endorsement, “consulting” and “management consulting,” under the same user. This dilutes the strength of an Endorsement as well as the credibility of the individual receiving it.

Through supervised training (a form of Machine Learning), we can identify a set of Endorsement categories (sub-populations), make valuable observations of interrelationship, and even classify unseen input objects.

For example, out of a sample size of 100 people that have been Endorsed for “Strategy,” criteria is set (wherein the threshold is set by analyzing the distribution of the total dataset) indicating that X # amount must be Endorsed for both “Strategy” and “Consulting” to be considered to have a parent-child or parent-parent relationship in that field of “Consulting.” If the frequency is met, then we can assign a relationship to “Strategy” and “Consulting.” The supervised learning algorithm will analyze similar examples as training data and produce an inferred function which can be used for mapping new inputs (BIBLIOGRAPHY #1). Through this Training Model, an Endorsement taxonomy is formed while the process is automated to accommodate any new Endorsements, any new taxonomies, and all potential parent-child or parent-parent relationships.

T-2: Taxonomy of Career Progression by Industry and Title: Using statistical classification to recognize patterns of Career Models

The variables (Titles aka “Manager” or “Financial Analyst” or “Partner”) themselves are in a Bernoulli distribution, and through supervised training (a form of Machine Learning), we can identify a set of Title categories by Industry (sub-populations) and make valuable observations of interrelationship. For example, the Management Consulting field is dominated by a 4-stage Career Model prior to Partner. Through examples such as the management consulting route, using input and desirable output (supervisory signal), a hierarchy and/or taxonomy for Career Models can be produced.

The “roadmap” of an individual on Linkedin in the context of Title and Industry may differ greatly from one to another (especially by industry and field), nevertheless, this pool of aggregate data (if analyzed through machine learning) can reveal traditional and non-traditional career paths that may be of value during Social Graph and Keyword Analysis, as well as to “visualize” career trajectory across any population. Literature has supported this method of taxonomy and finds it suitable for situations where there are huge amounts of unlabeled data and a small quantity of labeled data (BIBLIOGRAPHY #2).

Social Graph Analysis:

By means of applying hierarchical clustering, in this case, K-mode cluster, to LinkedIn Social Graph data, qualitative variables are put through an iterative refinement process by which dimensionality is reduced. Subsequently, a “Credibility Score” is constructed for Endorsers based on their specific career and/or professional characteristics.

For example: If Person A has 10 total endorsements for “Consulting,” he/she has a strength of 10 to endorse others for the same characteristics, factoring in sub-population data from Taxonomy Analysis (see above).

Furthermore: If he/she decides to endorse two individuals for the same category of Endorsement (e.g. “Consulting”), then each of them will have a strength of 5. This algorithm will involve analysis of node centrality (importance) with exploration of properties such as indegree (how many directed edges are incident on a node), outdegree (how many directed edges originate at a node) or degree (number of the previously mentioned incidents). We will also explore a variety of approaches to finding the end nodes of the Endorsement Graph and for developing a prediction algorithm for a target variable, in order to determine the degree of credibility of the endorsement – example approaches include: decision tree, rule induction, instance based methods, or a combination of the methods (BIBLIOGRAPH #3 AND #4).

An additional credibility factor will be determined by Career Model progress from Taxonomy Analysis. For example: High levels of Career Progress may be correlated to high value Endorsements. Similarly, High Value Employers (Government, For Profit, Non Profit, etc), and length of employment, may also be correlated to high value Endorsements (BIBLIOGRPAHY #5).

Keyword Analysis:

We will analyze any correlation of Endorsements (taking into account Taxonomy Analysis for Sub-Populations and Social Graph Analysis for Credibility) and Career Model (taking into Taxonomy Analysis of Titles as well as Employer, and Geography) compared to public data sources, for salary, time off, benefits, other perks, non-traditional compensation, etc.

We can find the “line of best fit,” coefficient of determination, R-squared value of correlation between an Endorsement (including Taxonomy Analysis and Social Graph Analysis), and Title (including Taxonomy Analysis), with Linear Regression Analysis. We can potentially observe distribution and even other characteristics like standard deviation and Z-score or T-Score. This allows us validate the Taxonomy Analysis and Social Graph Analysis if the observed R-squared is unsatisfactory, as well as building a predictive model that allows for forecasting outcomes.

Questions to Answer from these Analysis Data Models (Taxonomy, Social Graph, Keyword):

Are High-Credibility Endorsements (including Taxonomies) correlated with Career Model Success and High Salary, and therefore considerable as “High Value Skills”?

Is there a high frequency of High-Credibility Endorsements without Career Success?

Is there a high frequency of Career Success without High-Credibility endorsements?

Are certain Endorsements and/or Endorsers more valuable than others from a Career Model and/or Salary perspective?

Is Education a key driver for High Value Skills? What levels of Education? Which Educational Institutions?

What are other drivers for High Value Skills? For example: Internships, abroad experience, Emotional Quotient/Soft Skills, Experience with Diversity, Non-Profit Activity, Giving, etc.

Other Potential Research:
An Economic Analysis approach might be possible by analyzing the cost to achieve certain Career Models and/or Titles, such as Certification, Degree, Research, Tenure, etc., that might produce a Return on Investment model for achieving certain Endorsements by certain levels of Credibility, by comparing the Cost to known Salaries. This might produce a financial indicator of success by Endorsement based on Endorsement Credibility, all things considered.

What is the potential for utilizing High Value Skills to incorporate into Human Resources, Talent Model, and Educational Systems, to train the workforce?

Is it possible to generate Human Resources, Talent Model, and/or Educational Roadmaps based on High Value Skills, and/or determine Gaps by Population of High Value Skills to inform Upskilling and related investment?

A list of the data from LinkedIn that you believe you will need or want to access to conduct your proposed research

Endorsements, Geographies, Titles, Employers, Education Levels and Institutions, Certifications, Recommendations (including 360 degree understanding of the Recommendation) Key Words on profile, and other forms of Meta Data (rate of sign in, time on site, number of links, density of links)

Resources you plan to invest, e.g. how many people will join the research, what help you are looking for from LinkedIn side, and how much time you will need to complete the research

1 Full Time Resource: Data Science Administrator and Program Director – FTE from The

Tesla List and/or Tesla Foundation/National Police Athletic League

3.5 Part Time Resources: From the Northeastern University Coop Masters Students – The Tesla List has a strategic partnership with the Experience Network (XN) at NEU, with a focus on Data Analysis and Curriculum Development

Preferably: Cloud Computing Resources from LinkedIn, as well as Cloud Storage. Data should come either from API's or regular scrapes in flat file format (batched Daily/Weekly/Monthly)

Research should be completed within 90 days of receiving files or API access, with final presentation 30 days after research is complete. Total time: 120 Days.

Timeline, execution plan, key deliverables and milestones for the project

Total Program Time: 120 days

Research Period:

- Data Collection/Extraction and Initial Analysis: 15 Days
- Initial Data Model Development and Supervised Training: 30 Days
- Pilot Execution and Data Model Supervised Training Enhancements: 30 Days
- Final Data Model Execution and Insight Development: 15 Days

Milestone #1: Initial Presentation of Findings to LinkedIn and other Strategic Partners

- Initial Presentation Feedback Period: 15 Days
- Final Presentation and Report Development: 15 Days

Milestone #2: Final Presentation of Findings to LinkedIn and other Strategic Partners

Key Deliverables:

- Final Report and Presentations
- Data Model and Data Model Notes
- Video Documentary of our Efforts, including Tie-ins and Collaborations with Students form The Tesla List in Cleveland, Ohio (PAL and MC2)

Names, affiliations, postal addresses, phone numbers, email address, and LinkedIn profile of the participants

William Donovan, Founder and Interim Executive Director of The Tesla List

- Phone: 469-264-2903
- Address: The Tesla List C/O National Association of Police Athlectic/Activities Leagues, Inc. 12161 Ken Adams Way, Suite 110 RR, Wellington, Florida 33414
- https://www.linkedin.com/in/will0donovan/

Carma Khatib, Director and Founding Member, The Tesla List

- 703-992-3044
- Address: The Tesla List C/O National Association of Police Athlectic/Activities Leagues, Inc. 12161 Ken Adams Way, Suite 110 RR, Wellington, Florida 33414
- https://www.linkedin.com/in/carma-khatib-508a29115/

Lizi Feng, Co-Op Intern, Experience Network, Northeastern University, Masters in Data Analytics

- Phone: 510-386-8261
- Address: Northeastern University, 6024 Silver Creek Valley Rd, San Jose, CA 95138
- https://www.linkedin.com/in/lizi-feng-98097335/

Other participants from the Experiential Network (XN) at Northeastern University will be from the Masters Program in Data Analytics – potentially up to 5 Masters Candidates as a Capstone Project for the Masters Program

A short bio highlighting your background, expertise, and achievements, including prior relevant research

ORGANIZATION:

Our efforts are related to the Data Science Strategy of The Tesla List, a Non-Profit Strategic Partnership of the Tesla Foundation and the National Police Athletic League (PAL).

The Tesla List has developed a specific Talent Model and a Curriculum based on that Talent Model, which has been piloted as a Mentoring program at PAL Chapters in Cleveland, OH and Flint, MI.

The Curriculum is the official Entrepreneurialism Program at the MC2 Charter School at the General Electric Campus in Cleveland. By performing the research specified in this proposal, The Tesla List will be able to enhance its Curriculum's focus on High Value Skills – specifically skills correlated with High Performance Leadership, Creativity, EQ, Collaboration, and Innovation.

With 500 Chapters and 2,000,000 youth members, PAL represents an enormous opportunity to benefit under-privileged children and under-resourced communities with a disruptive educational system that utilizes Big Data and Analytics to drive positive outcomes.

BIOGRAPHIES:

William Donovan ran the Performance Management Program for 3 Global US Military Supply Chains in the United States, Europe, Central Asia and the Middle East. This included developing a comprehensive analytics capability that predicted sourcing, procurement, and on-time delivery performance across all categories of supply, utilizing data from approximately $500,000,000 in financial transactions and over $2,500,000,000 in sourcing data. William was a Subject Matter Expert on several Casandra-based Big Data projects for the Emirati Government in Dubai while he worked for Accenture, and is a top expert and consultant to the Pentagon on Big Data and Digital Supply Chain, in his role as a Program Lead on 2 military branches' ERP Transformations.

Carma Khatib is a seasoned operator in Technology Project Management and has worked to develop the Data Science Strategy for The Telsa List.

Lizi Feng performed data analytics functions for both the public and private sectors for 3+ years, with clients including the World Bank, United Nations Foundation, UCSF Medical Center, and the U.S Environmental Protection Agency. She is currently a healthcare consultant to growth technology companies in the Bay Area valuating at as much as 28 billion. Her pilot financial literacy program won the Rudd Foundation award in 2014.
Some of her relevant research can be found here: http://berkeleyair.com/wp-content/publications/SPT_Inventory_Report_v3_0.pdf.

Coding language (e.g. Python, R, Pig/Hive) skills and experiences

Python – Advanced knowledge of statistical tools and other research and development functions

R – Advanced knowledge of statistical analysis and forecasting, proficiency in data wrangling and data manipulation for predictive modeling and other supporting packages

SQL – Advanced knowledge of constructing relational database, data controls, data manipulation, and data extraction

Orange – Advanced application of data mining and empirical evaluation of machine learning algorithms; and its use as a Python library

Additional details

The Tesla Foundation is a non-for-profit science and technology Think Tank focused on the transition from the Industrial Revolution 3.0 to 4.0 and the Architect of the Tesla Technology Farm System. As we navigate the shift from the “Information Age” to the “Autonomous Age,” education and workforce development must evolve quickly and with great urgency to meet the demands of the “New Economy.” The Tesla Foundation accomplishes its non-profit goals utilizing the combined efforts of its technology farm system, research, education, public private partnerships, applications, and high level educational events and summits.

The National Association of Police Athletic/Activities League (National PAL) and its chapters work nationwide promoting the prevention of juvenile crime and violence by building relationships among kids, cops and community through positive engagement.

Bibliography

1. Kotsiantis S. B. “Supervised Machine Learning: A Review of Classification Techniques”. https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf
2. Triguero, Isaac. Garcia, Salvador. Herrera, Francisco. “Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study”. https://pdfs.semanticscholar.org/afbf/0b849011fe3813bb74d9a17157b026f49a8a.pdf
3. Langley, Pat. “Application of Machine Learning and Rule Induction”. Retrieved from:http://scholar.google.com/scholar_url?url=http://www.dtic.mil/cgi-bin/GetTRDoc%3FAD%3DADA292607&hl=en&sa=X&scisig=AAGBfm0T3hY_Ai8-wvxfGRJQ7ws8VVcqeQ&nossl=1&oi=scholarr
4. Baumgartner, Peter. “Instance Based Methods” September 2005. Retrived from: http://users.cecs.anu.edu.au/~baumgart/slides/tableaux-2005-tutorial/handout.pdf
5. Warta, David. “Personal Foul: Unnecessary Restriction of Endorsement and Employment Opportunities for NCAA Student – Athletes”. Retrieved from: http://digitalcommons.law.utulsa.edu/cgi/viewcontent.cgi?article=2445&context=tlr

Please send your proposals in PDF format to EconomicGraphResearchProposal@linkedin.com with the following subject line: “Analytics/Economics/Relevance: [Proposal title]” by June 1, 2017. By downloading and submitting a proposal, the individual submitting the proposal and each teammate agree they have read, understood and agree to be bound by the proposal requirements on the Economic Graph Research website. The individual submitting the proposal and each teammate agree that LinkedIn may conduct use, research, or develop software, products, or features that are the same or similar to your proposal and that you waive any claims with respect to any use, research, or developments by LinkedIn.