Principle Investigator's Social Networks Effect on Project Success

Adam Ryman, Antonio Sanfilippo, Eric Bell

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

Purpose: Improve Research Output

Determining the optimal way to execute a research project is a primary goal for both research institutions and funding agencies. Here, using web harvesting and social analytics, a social network was both built and analyzed using data from internal PNNL projects and Google Scholar's database of research papers. Results from this analysis now can be used to help model an optimum research project situation.
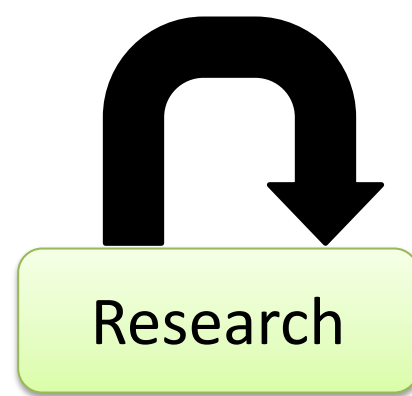
Research

Model: What has been done?

The Knowledge, Discovery, and Informatics group of the Computational and Analytics division of PNNL has created a model that takes attributes of a research project (such as funding, number of publications, aspects of the people working on the project, etc.) and optimizes to any of said attributes.

This model is processed and analyzed by machine learning algorithms in a program known as Weka.

The output of this model can help improve the research process by allowing research and funding organizations to understand the optimum way to run a research project.

Research Question

Funding — Publications

People — Job Creation

Overall Project Success
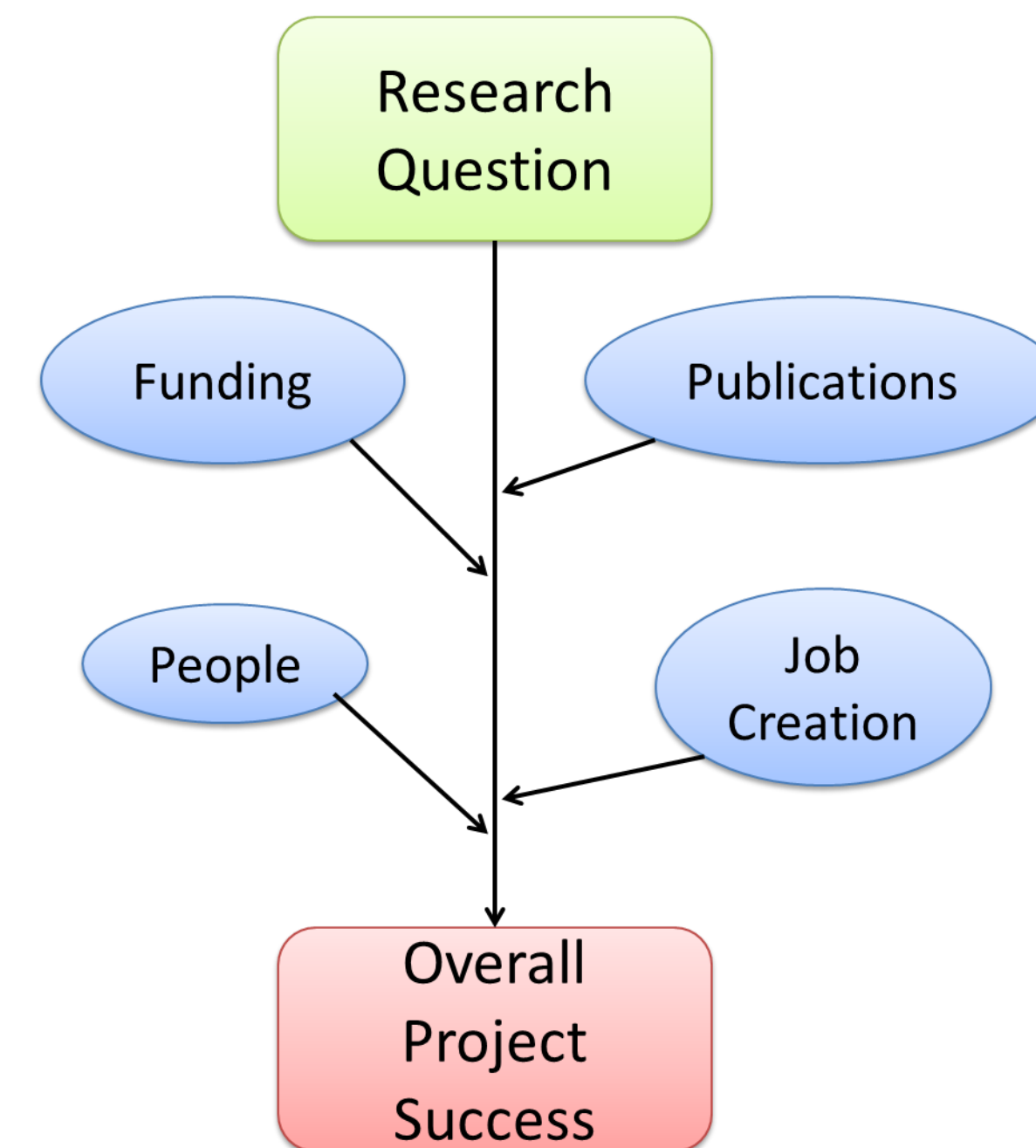
Specific Goal: To add social network related attributes to our current model

Currently, our model has over 200 attributes. The goal of this project is to create addition attributes in our model in order to increase its precision and accuracy.

These additional attributes, as specified by Antonio Sanfilippo, describe Principal Investigator's social networks strengths and weaknesses. These kind of attributes may possibly have a huge impact on our model as a strong social network is akin to having a strong access to resources.
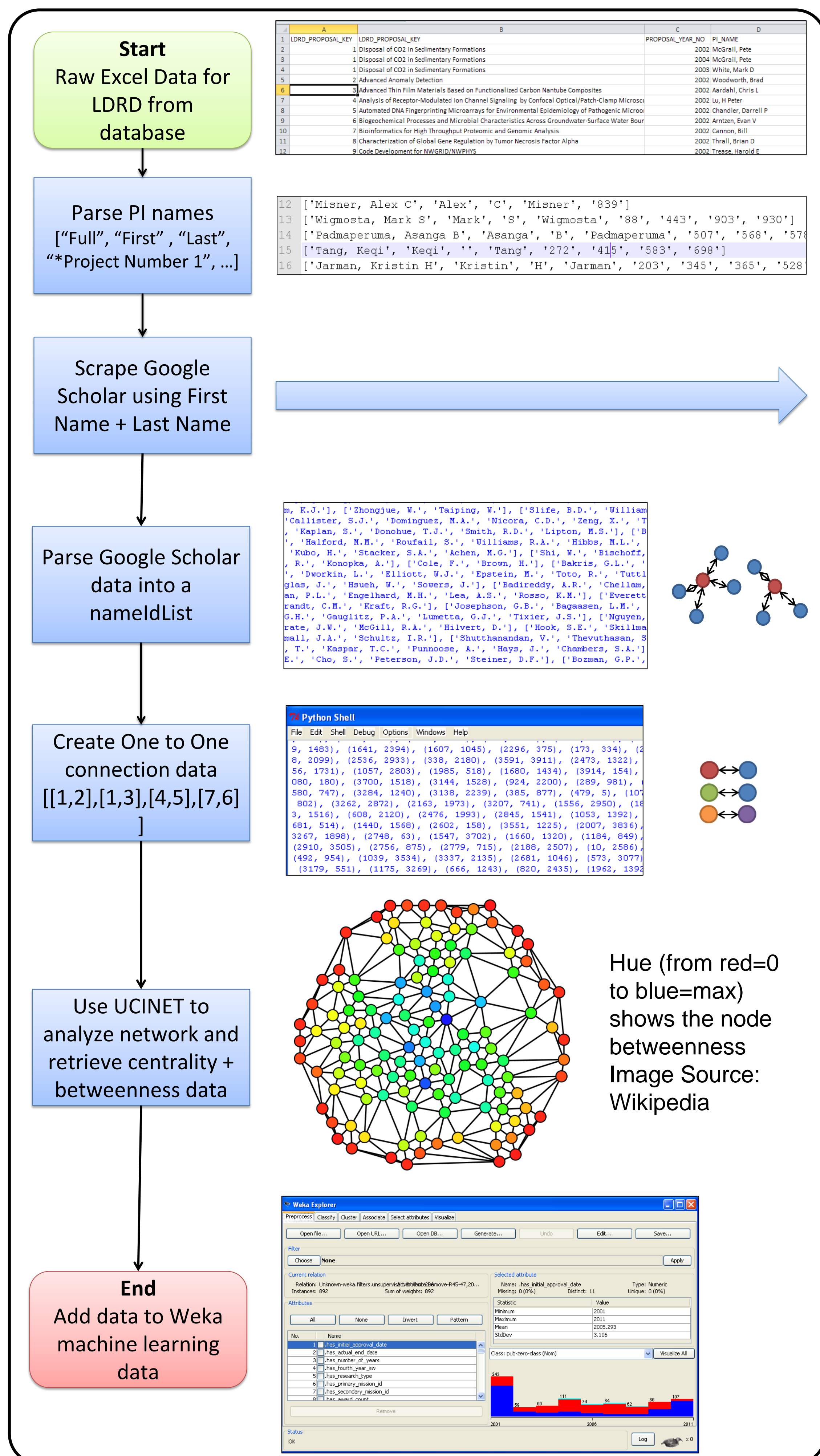
Question:

Does a Principal Investigator's social network correlate in any way with the success of the projects they manage?
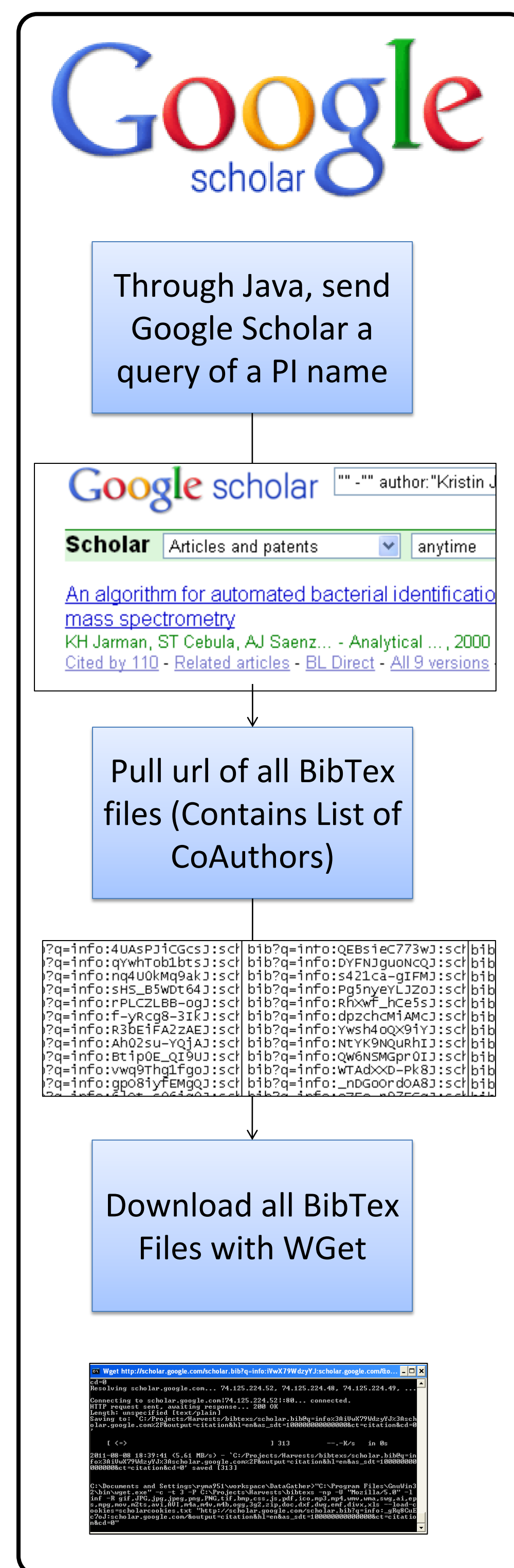
Goals:

1. Construct a social network for all PI's using co-authorship In publications.
2. Analyze social network for features.
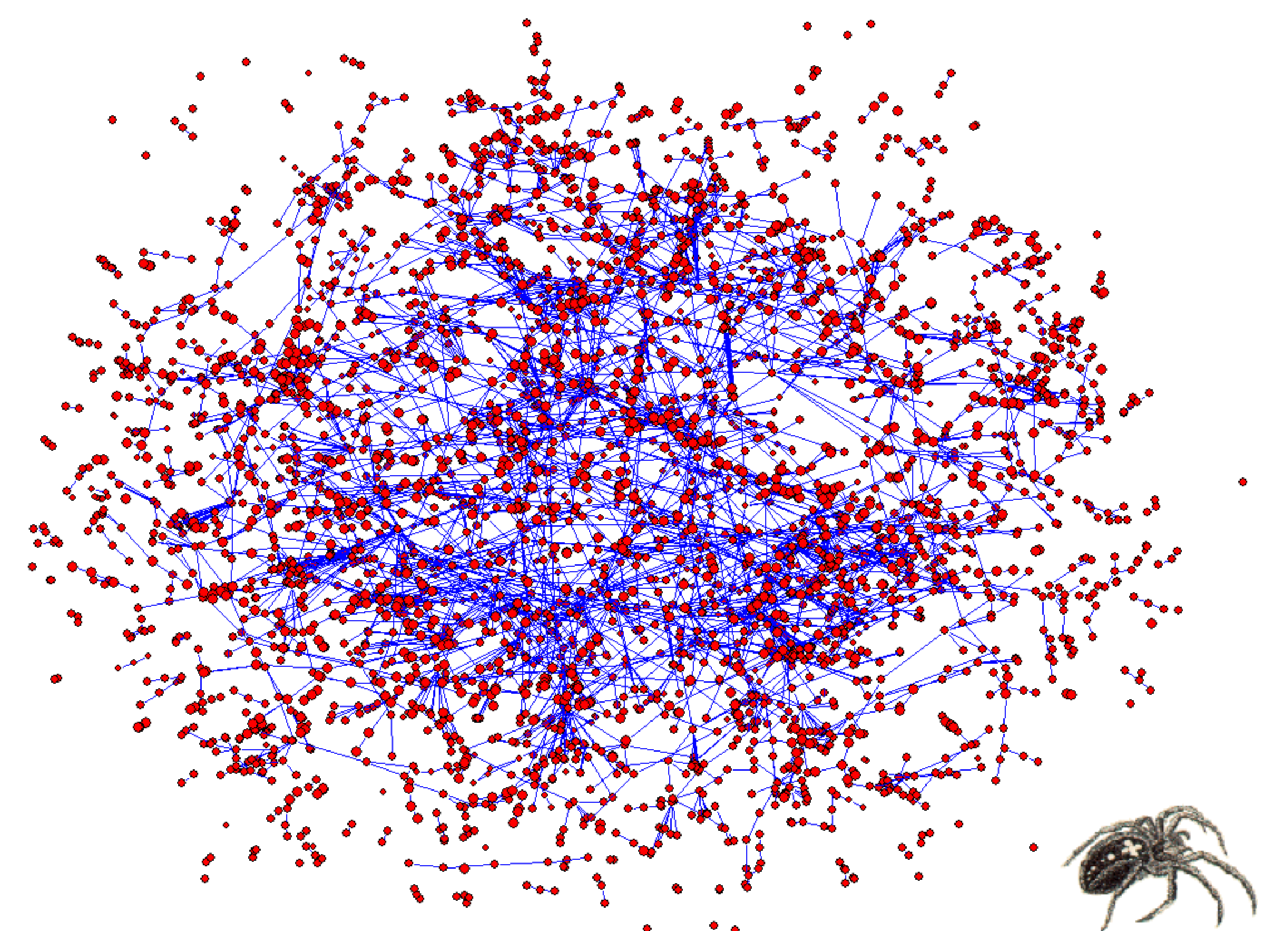3. Map features to existing model.

Co-Author Network Visualized in Pajek

Method / Workflow

Start
Raw Excel Data for LDRD from database

Parse PI names
["Full", "First", "Last", "*Project Number 1", …]

Scrape Google Scholar using First Name + Last Name

Parse Google Scholar data into a nameIdList

Create One to One connection data
[[1,2],[1,3],[4,5],[7,6]]

Use UCINET to analyze network and retrieve centrality + betweenness data

End
Add data to Weka machine learning data

Hue (from red=0 to blue=max) shows the node betweenness Image Source: Wikipedia

Google Scholar Flow

Through Java, send Google Scholar a query of a PI name

Pull url of all BibTex files (Contains List of CoAuthors)

Download all BibTex Files with WGet

Current Progress

Currently, data is still being harvested though Google Scholar.

As shown in the image to the right, after a large amount of queries, Google locks access and requires a CAPTCHA which the current automation software is not able to handle. A delay of 4 – 6 hours must elapse and a manual check for lost data must occur before data harvesting restarts.

After our data is fully harvested, all that is required is that it is pushed down the rest of our analyzing workflow. And whether or not the results of our processing shows correlation, it will never the less improve our understanding of attributes that lead to project success.

Software / Programming Languages

- Java
- Eclipse
- Wget
- Python
- IDLE
- Weka
- Pajek
- UCINET
- NotePad++

Acknowledgments

Antonio Sanfilippo: Mentor and Project Leader

Eric Bell: Advisor, Google Scholar information

DOE / PNNL : Funding and Intern Program

Contact

Adam Ryman
Phone : 425-289-9391
Work Email : adam.ryman@pnnl.gov
Home Email : adamryman@gmail.com