Automatic Citation Extraction and Analysis
Tyler J. Conn, Samuel Hill, Ahmed Mustafa, Dr. Andrea Tartaro

Introduction
Large scale scientific research projects can result in breakdowns in communication, planning, and eventually goals. To promote and measure cohesion within a project, we propose analyzing bibliographic cross citing and textual similarities in research papers. To decrease effort required from the research scientists, this information must be automatically extracted from the research papers. This information can then be used to identify cohesion in the research project and possibly suggest new collaborations.

DOI Extraction
A DOI, or Digital Object Identifier, is a unique string of letters and numbers used to identify a single piece of published work online. By automatically extracting a DOI from a research paper, we can then use it to identify the article within the PubMed database and extract its information for later use.

Information Extraction
After obtaining the DOI, it is entered into the MetaPub module to extract information on both the original article and all of its citations [1]. The pulled information ranges from author names to dates or subjects in order to find the most possible connections. This data is then transferred to a CSV that is read by a python script that creates the network of information in a RDF(directed, labeled graph data format in the Web) database called SPARQL. The data is then analyzed through the Weka data mining software [2].

Future Work
The current work is a proof of the concept. For the future, there are several important aspects to build upon is the size of the database,. With more data to work with, the algorithm could find countless more connections amongst the scientific community. Further implementation could compare author, subject, and even publisher connections. Eventually the use of learning algorithms such as Neural Nets and Support Vector Machines could help to identify groups of similar papers and classify text to find similarities on the textual level [3]. These tools could then be used to report on collaboration amongst groups and to suggest collaborations similar to Spotify but for research [4].

Sources
[1] metapub 0.3.17.1; https://pypi.python.org/pypi/metapub/0.3.17.1

Acknowledgements
This research is based upon work supported by the National Science Foundation-EPSCoR program under Grant Number EPS-0903795. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.