**Integrative Research Collaboration Pipeline in Expertise Classification and Researcher Network Visualization**

**Background and Aim**

Research collaboration (RC) is broadly defined as the working together of researchers to achieve the common goal of scientific knowledge production (Katz and Martin, 1997). This collaborative effort can occur at the institutional level, between partnered institutions or internationally, and reflects the scientific collaborative networks in which modern science and researchers are embedded (Fonseca et al, 2016). RC is a significant outcome within university institutions as a target for innovation through shared expertise, ideas and resources, reducing university expenditure and ultimately increasing academic productivity in the generation of new scientific knowledge (Fonseca et al, 2016) (Lee & Bozeman, 2005). With the rise of the electronic age, medical and scientific databases such as PubMed, Web of Science and Scopus, in addition to researcher profile databases such as ORCID, now provide pervasive online databases capturing these collaborative relationships between researchers, serving as the ‘webbing’ for RC networks represented by academic relationships such as co-authorship. This provides an opportunity for research organisations to utilise collaboration data in the application of research network visualization, analysis and prediction or recommendation of future researcher collaborations, in an effort to expand and promote research collaboration, expertise exchange and research diversification amongst these organisations through a shift towards more ‘active’ data-driven solutions in the support of institutional research collaboration (Falagas et al, 2008)(Lee & Bozeman, 2005)(Fonseca et al, 2016).

There is an abundance of research subfields involved within the domain of RC networks, particularly in the form of expertise classification and recommendation, RC network visualisation and link prediction (Yan et al, 2018) (Sun et al, 2015)(Deng et al, 2008) (Zhou et al, 2019)(Sun et al, 2011) (Miller et al, 2014) (Silvia et al, 2013) (Afzal & Maurer, 2011) (Fagan et al, 2018) (Kang & Coppel, 2015) (\*Sun et al, 2011) (Luo et al, 2014). However, there appears to be a lack in the provision of a pragmatic pipeline and integrated delivery for university institutions. This paper aims to fill this university resource gap, proposing an integrated pipeline architecture and RC prototype-tool in the utilisation of ‘static’ university research databases and complementary online scientific databases in shifting this research data source to an ‘active’ institutional role through the integration of these segmented methodologies, filling the lack of university data-driven solutions available in the promotion of research collaboration amongst university researchers and their institutional partners.

A close up of text on a white background

Description automatically generated

**Design, Methodology and Timeframe**

This proposed project will consist of *four technical stages* of six total stages based on the complete pipeline framework of figure 1, spanning over a *three-semester duration* and with the fifth and sixth stage outlining future trajectories for the tool. The overall project is intended as a prototype or proof-of-concept for future implementations following figure. 1. However, it is expected *only stages one to four will be completed* within this dissertation project, with stage five and six only providing future context for the project within the overall pipeline framework.

***Stage 1: Data Extraction***

The first stage sets the foundation for the collaboration tool, *data extraction*. This involves two key processes; the raw data extraction from scientific and medical databases, namely PubMed, Scopus and Web-of-Science, primarily through utilisation of database unique identifiers and secondarily ORCID IDs, with these being derived from research profiles within the University Researcher Database (URD), in this case SPHERE, provided by UNSW. Using these researcher profile datasets consisting of all researchers published papers (titles, abstracts, published journal or conference and digital object identifiers) , co-authors and university affiliations for each URD researcher from each respective scientific/medical database will be extracted for stage two, with variation in paper coverage expected between these databases as a consequence of the specific purposes of each database (Falagas et al, 2008) (Luo et al, 2014).

***Stage 2 and 3: Author Disambiguation***

Stages two and three aim to be complementary solutions to the author disambiguation problem inherent within researcher database studies and analysis (Xu et al, 2018). This problem arises as an author cannot necessarily be uniquely identified by his or her name when extracted from online researcher databases and additionally a researcher’s papers extracted may not necessarily all belong to the researcher in question. *Stage two* firstly solves this problem of ‘*right papers’* through a network embedding based method for author disambiguation offered by Xu et al (2018) (Xu et al, 2018). Using the researcher data extracted in stage one, five networks for each ambiguous researcher set will be constructed using five distinct relations (co-title, co-abstract, co-affiliation, co-venue and co-author), network embedding will then be undertaken to encode the global information of these network graphs into paper representations. Finally, using a clustering algorithm these paper representations will be partitioned into paper clusters, with each containing all papers for a unique author (Xu et al, 2018).

*Stage three* aims to distinguish the most likely correct author (researcher) cluster match to the URD researcher using network comparison, comparing between respective researcher paper clusters and a multi-relational network construction of the researcher drawn directly from the university researcher database researcher and publication data. As networks are almost never identical, inexact graph matching through Unknown node-correspondence (UNC) methods based on global statistics such as the clustering coefficient will be used to ascertain the extent of similarity between researcher clusters to the reference UNC network, having the property of converging to zero as networks become isomorphic or identical (Leva et al, 2018). Thereby, allowing for the most similar researcher cluster to be identified as most representative of URD researchers research profiles, ensuring the ‘*right researcher’* has been extracted. Finally, in concluding extraction and disambiguation these separate disambiguated researcher datasets will be merged into a final cleaned researcher dataset and a manual validation procedure will be undertaken to ascertain the accuracy in matching URD researcher profiles with extracted research profiles based on a small subset of UNSW researchers (n) representing the known ‘ground truth’ and serving as the final validation process for data extraction and disambiguation and providing a data fixture for assessing correct output and functionality in future runs of the application.

***Stage 4: Researcher Expertise Classification***

Stage four is concerned with the representation of researcher expertise, through the classification of researcher topics in paper sets merged and disambiguated in stage one to three. This will be undertaken through the text mining scheme approach, applying a Latent Dirichlet allocation (LDA) topic-model derived from the R’s ‘topicmodels’ package over researcher paper sets, summarising topic vectors for a researchers paper set of abstracts and titles as a summative topic centroid representing each researchers topic distribution according to topic distributions extracted from training on all available researchers papers (titles and abstracts) (R Core Team, 2017) (Katsurai et al, 2016). This is a more traditional methodology and fitting for the time restraints of this project as compared with more semantically inclusive machine learning approaches (Yan et al, 2018). This final stage of the project is intended to be the end-point of the dissertation, serving as the foundation for functionality in allowing expertise representation to be used as a feature for the visualization of institutional expertise and within collaboration recommendations in future stages of the pipeline\* (Deng et al, 2008) (Kang & Coppel, 2015) (Afzal & Maurer, 2011) (Katsurai et al, 2016).

*\*See* ***appendix*** *for detailed account of stages five and six, which are not expected to be undertaken within this project.*

**References**

Afzal, M. and Maurer, H. (2011). Expertise Recommender System for Scientific Community. *Journal of Universal Computer Science,* 17(11), pp. 1529-1549.

Bastian, M., Heymann, S,. & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the Third International Conference on Weblogs and Social Media.

Deng, H., King, I. and Lyu, M. (2008). Formal Models for Expert Finding on DBLP Bibliography Data. In: *Proceedings of the 8th International Conference on Data Mining*. Pisa: IEEE Computer Society, pp. 163-172.

Du, Y., Wang, C. and Ji, J. (2017). Biomedical semantic indexing by deep neural network with multi-task learning. In: *IEEE International Conference on Bioinformatics and Biomedicine*. Kansas City: BMC Bioinformatic, 20(502), https://doi.org/10.1186/s12859-018-2534-2.

Fagan, J., Dolly, J., Vanderford, N. Weiss, H. and Levens, J. (2018). Assessing Research Collaboration through Co-authorship Network Analysis. J Res Adm. 49(1), pp. 76-99.

Falagas, M., Pitsouni, E., Malietzis, G. and Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *The FASEB Journal*, 22(2), pp. 338-42.

Fonseca, B., Sampaio, R., Fonseca, A. and Zicker, F. (2016). Co-authorship network analysis in health research: method and potential use. *Health Research Policy and Systems, 14(34),* https://doi.org/10.1186/s12961-016-0104-5.

Kang, Y., Li, Y. and Coppel, R. (2015). Capturing Researcher Expertise through MeSH Classification. In: *Proceedings of the 8th International Conference on Knowledge Capture*. K-CAP, pp. 1-8.

Katz, J. and Martin, B. (1997). What is research collaboration?. *Research Policy*, 26(1), pp. 1-18.

Katsurai, M., Ohmukai, I. and Takeda, H. (2016). Topic Representations of Researchers’ Interests in a Large-Scale Academic Database and Its Application to Author Disambiguation. *IEICE Transactions on Information and Systems.* 9(4), pp. 1010-1018.

Lee, S. and Bozeman, B. (2005). The Impact of Research Collaboration on Scientific Productivity. *Social Studies of Science,* 35(5), pp. 673-702.

Li, J., Zhao, D., Ge, B., Yang, K. and Chen, Y. (2018). A link prediction method for heterogeneous networks based on BP neural network. *Physica A,* 495(C), pp. 1-17.

Liben-Nowell, D. and Kleinberg, J. (2007). The Link-Prediction Problem for Social Networks. *Journal for the American Society for Information Science and Technology,* 58(7), pp. 1019-1031.

Luo, j., Pelfrey, C. and Zhang, G. (2014). Visualizing and Evaluating the Growth of Multi-Institutional Collaboration Based on Research Network Analysis. In: *AMIA Joint Summits on Translational Science proceedings.* AMIA Joint Summits on Translational Science, pp. 60-66.

Miller, L., Gazan, R. and Still, S. (2014). Unsupervised Classification and Visualization of Unstructured Text for the Support of Interdisciplinary Collaboration. In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. CSCW, pp. 1033-1042.

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rossum, V., Guido., and Drake, F. (1995). Python tutorial. *Centrum voor Wiskunde en Informatica Amsterdam*, The Netherlands.

Silva, T., Guo, Z., Jiang, H. and Chen, H. (2013). A social network-empowered research analytics framework for projection selection. *Decision Support Systems*, 55(4), pp. 957-968.

Shakibian, H., Charkari, N. and Jalili, S. (2016). A multilayered approach for link prediction in heterogeneous complex networks. *Journal of Computational Science*, 17, pp. 73-82.

Shakibian, H. and Charkari, N. (2018). Statistical similarity measures for link prediction in heterogeneous complex networks. *Physica A*, 501, pp. 248-263.

\*Sun, Y., Barber, R., Gupta, M., Aggarwal, C. and Han, J. (2011). Co-author Relationship Prediction in Heterogeneous Bibliographic Networks. In: *International Conference on Advances in Social Networks Analysis and Mining.* Kaohsiung: IEEE Computer Society, pp. 121-128.

\*\*Sun, Y., Han, J., Yu, P. and Wu, T. (2011). PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. In: *Proceedings on Very Large Data Bases (VLDB).*

Sun, J., Xu, W., Ma, J. and Sun, j. (2015). Leverage RAF to find domain experts on research social network services: A big data analytics methodology with MapReduce framework. *Int. J. Production Economics*. 165(5968), pp. 185-193.

Tantardini, M., Leva, F., Tajoli, L. & Piccardi, C. (2019). Comparing methods for comparing networks. *Sci Rep.* 9(1), 17557.

Xu, J., Shen, S., Li, D, and Fu, Y. (2018). A Network-embedding Based Method for Author Disambiguation. In: *Proceedings of the 27th ACM International Conference of Information and Knowledge Management.* Torino: CIKM, pp. 1735-1738.

Yan, Y., Yin, X., Yang, C., Li, Sujian, Li. and Zhang, B. (2018). Biomedical Literature classification with a CNNs-based hybrid learning network. *PLOS ONE, 13(7): https://doi.org/10.1371/jounral.pone.0197933*

Zhou, H., Sun, J., Zhao, Z., Yang, Y., Xie, A. and Chiclana, F. (2019). Attention-Based Deep Learning Model for Predicting Collaborations Between Different Research Affiliations. *IEEE Access*. 7, pp. 118068-118076.

***Appendix***

***Stage 5: Network Visualization***

Using Visualization software available within R’s Viznetwork or ggnet2 packages in addition to Gephi’s Network Visualization software, numerous visualizations between homogeneous research networks will be undertaken to find the most insightful and useful visual analysis of the researcher institutions network through a range of ‘lens’ available within the extracted and cleaned data (Bastian et al, 2009). These being; co-topic network visualization to assess the distribution of expertise, co-author network to assess the distribution of research collaboration amongst researchers, co-affiliation network visualisation to assess relationships between organisations and organisation collaboration and co-venue network visualisation to assess the distribution of institutional research amongst academic journals and conferences.

***Stage 6: Researcher Collaboration Recommendation***

The final stage combines both dataset outputs from stage one to four applying these as relationship types (links) in the form of discrete meta-paths in research collaborative networks between researchers (nodes) to obtain a heterogenous multi-relational layered network representation of the data, with each path between researcher nodes carrying unique semantic dependencies or meanings between target researchers, as opposed to a less accurate homogeneous (single relation-type or layer) co-author network representation (Shakibian & Charkari, 2016) (Shakibian & Charkari, 2018). This meta-path approach allows for indirect relations to be captured between researchers through a combination of relational pathways in the research collaboration network; co-authorship through papers, authors, variations in expertise (topics), venues and/or affiliations. Similarity measures based on co-occurrence events along these meta-paths amongst all researcher pairs can then be calculated as a matrix representing these semantic network pathways – being analogous to topological features in a traditional dataset (Li et al, 2018). A novel link prediction machine-learning approach aimed at predicting researcher pair-wise collaborations in the future can then be applied to these extracted network statistics (\*\*Sun et al, 2011) (Shakibian & Charkari, 2018) (Liben-Nowell & Kleinberg, 2007) (\*Sun et al, 2011). R language will be used in the computation of similarity statistical measures, with Python then being used to train, test and evaluate a 3-layer deep neural network meta-path based prediction model, implemented through the ‘Keras’ library to predict which researcher pairs based on these relations; expertise (topic), author, paper and affiliation may be most likely to collaborate in the future, providing a data-driven architecture in supporting university RC decisions (Li et al, 2018) (Rossum et al, 1995)**.**