THE RESEARCHER SOCIAL NETWORK:

A SOCIAL NETWORK BASED ON METADATA OF SCIENTIFIC PUBLICATIONS

Yang Yang, Ching-man Au Yeung, Mark J. Weal and Hugh C. Davis

Contact: yy4@ecs.soton.ac.uk

1. Introduction

Scientific journals can capture a scholar's research career. **We hypothesize that data, such as, i) publication keywords, ii) personal interests, iii)the themes of the conferences where papers are published, and iv)co-authors of the papers; either directly or indirectly represent the authors' research interests, and by measuring the similarity between these data we are able to construct a researcher social network.**

Based on the four types of data, social network graphs were plotted, studied and analyzed. Interestingly, our results showed that the graph based on publication's keywords were more representative than the one based on publication's co-authorship (approach adopted by Newman). The findings from the evaluation were used to propose a dynamic social network data model.

2. Constructing A Researcher Social Network

Data Mining

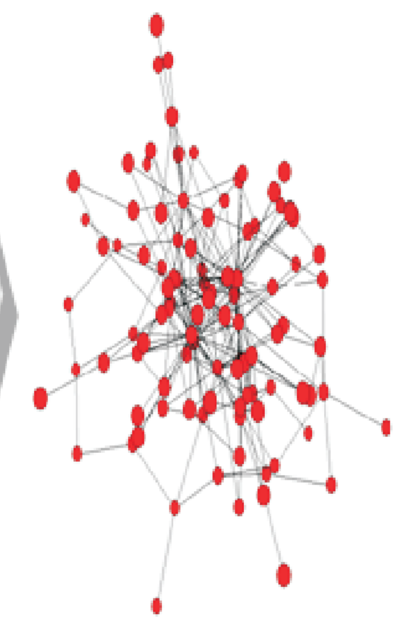
- **Data Source**
  - Personal Interests data obtained from Researcher Online profile RDF files
  - Keywords, co-authorship and the conference data are collected from **7,628** Eprints publications
- **Basic Intelligent inferring Processing**

Construct Network Graph

- **Tripartite Graph**
  - Researchers {R}
  - Keywords {K}
  - Publications {P}
  - Tripartite graph $T \subseteq R \times K \times P$.
- **Hypergraph**
  - $H(T) = \langle V, E \rangle$
  - Vertex $= R \cup K \cup P$
  - Edge $= \{(r, k, p) \mid (r, k, p) \in T\}$

Similarity Measuring

- **Measures the distance between two sets of data, taking into account the weight of each keywords**



3.Researcher Social Network Graph

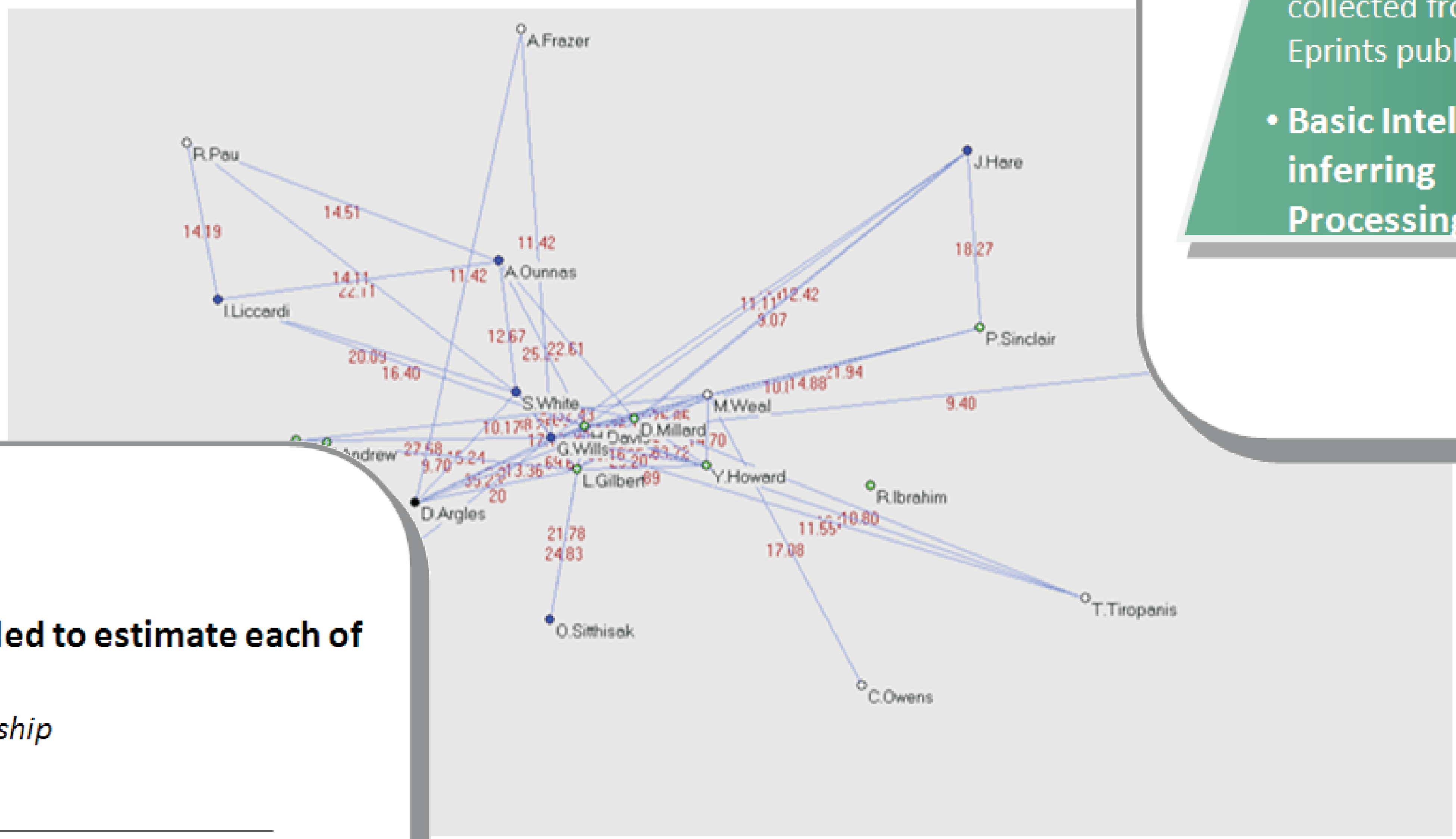


Figure 1: Researcher Social Network Graph based on blended data, where *Threshold* > 8, Keyword=0.32, Conference=0.35, Interest=0.13, and co-author=0.2.

4. Evaluation Results

Goal : In order to generate one "ideal" graph, we needed to estimate each of the data types' weight to blend all the data together.

$Edge_R = \alpha Keywords + \beta Interest + \gamma ConferenceData + \delta Co\text{-}authorship$

Table 1 – Rate of accuracy of each graph

	Graph			
	Keywords	Interests	Conference	Co-author
Connected People Circled	58.9%	51.9%	57.5%	71.5%
Connected People Crossed	20.5%	7.4%	21.8%	18.0%
Unknown	20.6%	40.7%	20.7%	10.5%

Table 2 – Rating of importance of data in terms of how it represent a scholar's research interests

Evidence	Estimation of weighting coefficient
Selected Keywords / Selected Interests	$\alpha \approx 2.5\,\beta$
Ranking of Representative of graph	$\gamma > \alpha > \delta > \beta$
Ranking of Importance of Data	$\alpha > \delta > \beta > \gamma$

5. Conclusions

- Using keywords to construct a social network was considered to have a better representativeness than using the co-authorship approach.
- The conference data has high noise, and was considered the least important data to construct a social network. However, after the conference data was visualized as a graph, it was considered to be the most representative amongst the others. The cause of this result can be attributed to the fact that the theme of the conference data is collected within a small research organization with a limited number of publications.
- We propose a dynamic social network data model:

$Edge_{social\ network} = X \times Edge_{personal\ attributes} + Y \times Edge_{social\ relations}$

- We believe that setting a higher weight for the personal attributes dataset (X) than the social relations (Y) will produce a better scientific collaboration social network which better captures the similarity in research interests.