The Researcher Social Network: A Social Network Based on Metadata of Scientific Publications

Yang Yang[1], Ching-man Au Yeung[2] , Mark J. Weal[1] and Hugh C. Davis[1]
[1]Learning Societies Lab, [2]Intelligence, Agents, Multimedia Group
School of Electronics and Computer Science, University of Southampton
Southampton, SO17 1BJ, UK
{yy4/mjw/hcd/cmay06r@ecs.soton.ac.uk}

ABSTRACT

Scientific journals can capture a scholar's research career. A researcher's publication data often reflects his/her research interests and their social relations. It is demonstrated that scientist collaboration networks can be constructed based on co-authorship data from journal papers. The problem with such a network is that researchers are limited within their professional social network. This work proposes the idea of constructing a researcher's social network based on data harvested from metadata of scientific publications and personal online profiles. We hypothesize that data, such as, publication keywords, personal interests, the themes of the conferences where papers are published, and 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 mentioned above, social network graphs were plotted, studied and analyzed. These graphs were then evaluated by the researchers themselves by giving ratings. Based on this evaluation, we estimated the weight for each type of data, in order to blend all data together to construct one ideal researcher's social network. Interestingly, our results showed that a graph based on publication's keywords were more representative than the one based on publication's co-authorship. The findings from the evaluation were used to propose a dynamic social network data model.

Keywords

social network, scientist collaboration network,

1. INTRODUCTION

From a researcher's profile publication list, it is possible to know the research interests of the researcher, his/her collaborators, and even the latest conferences attended by him/her. Harrison & Stephen [1] described the electronic journal as the heart of an online scholarly community, where academic journals function principally as channels of communication for practicing scholars. Newman [2] constructed a scientist collaboration network based on co-authorship data from journal papers. In this collaboration network, two scientists were considered connected if they had authored a paper together. Many of these networks' properties have been studied, but the benefits of such networks are rarely questioned. It is possible to develop a social network application based on this data model, and researchers are able to use such application to maintain a social relationship with their colleagues (the co-authors) or to explore who else their colleagues have been publishing papers with. However, the collaboration is normally established based on similar research interests. This raises a question, "If people publish a paper together, does it imply that they share similar research interests?" While it is arguable, we are convinced that this is not always the case – especially when a publication has a long list of authors, where each author may come from different research areas and is only partially involved in the project. Another problem is that researchers are limited to their own professional social environment in this network, which makes cross-field collaboration (through the network) almost impossible. Therefore, although the co-authorship network captures the researchers' professional social relations, the value of such network is very limited. Theoretically, if we could construct a social network which is based explicitly on the researcher's research interests, this network would be considered more useful than the simpler co-authorship network. In this network, two scientists will be connected if they have common research interests, and the more interests they share the stronger their connections will be.

In this paper, we propose to construct a researcher social network based on data extracted from personal online profiles and metadata of scientific publications, which could be employed to represent the researchers' interests. In particular, the data we investigate are: (1) keywords of publications; (2) personal interests; (3) data about conferences where papers were published; and (4) co-authors of publications. We hypothesize that these data can either explicitly or implicitly represent a researcher's interests, and by measuring the similarity between these data we are able to construct a researcher social network. The Learning Society Lab (LSL) of the School of Electronics and Computer Science at the University of Southampton was chosen to be the research target, as this is a coherent group which can be used as a research organisation structure model to compare with the computer generated social network. A total of 7,682 scientific publications from the EPrints publication repository and 63 on-line profiles from the LSL were studied and analyzed.[1] An experiment was designed to evaluate the quality of the researcher social networks and to estimate the weighting of each data type in order to blend these data together to construct an ideal social network – a network which represents all individual researchers' research focus, as well as capturing their social relationships.

[1] http://eprints.ecs.soton.ac.uk/

2. BACKGROUND

The traditional way to construct a social network of ties is by interviewing participants or by circulating questionnaires [3]. Today, the data about a community can be collected from various sources online. For example, a study about a network of movie actors [4]. The movies in which they appear have been compiled using the resources of the Internet Movie Database[2], which contains the name of nearly half a million actors. The root of the idea to construct scientific collaboration network can be found in Erdős number project[3] - Paul Erdős is the one person having an Erdős number of 0, researchers who had published a paper with Erdős were given an Erdős number of 1, those who had published with one of those people but not with Erdős, a number of 2 and so forth. Newman [2, 5] studied the idea of constructing a scientific collaboration network where two scientists are considered connected if they have authored a paper together. The construction of a collaboration network is straightforward; the vertices of the network are author names extracted from publications, and the edges are added between each pair of authors on each paper. However, the results indicate that simple co-author networks cannot capture the variation in the strength of collaborative ties. There is also research that studies large-scale networks representing research in mathematics, biology, physics, computer science and neuroscience [6-8]. Matsuo et al. [9] studied mining associations between conference participants from the Web. The authors targeted participants from a Japanese research conference and measured relationships between participants, based on search engine query results. The links were extracted by comparing the query results with the pre-set keywords; the relationship was extracted as "*co-author*", "*lab*", "*proj*", "*conf*". From the preliminary evaluation results, the author concluded that the "co-author" and the "lab" relation produced a high precision (96.2% and 87.0%) in terms of capturing actual social relations, whereas the "*proj*" and the "*conf*" produce a low precision indication (12.5% and 42.1%).

3. CONSTRUCTING ASOCIAL NETWORK

To construct a co-authorship network is straightforward, as co-authorship is explicitly stated in publication reference data. In social network analysis literature the co-authorship network is also known as an affiliation network, in which people are connected by edges with weights corresponding to the strengths of their affiliations. However, using simply the authors' names to construct a network can be problematic because the name of an author can be written in several different formats such as in full or with initials. In the EPrints dataset we collected, each author is represented by a unique identifier. Therefore, the co-authorship network we constructed by using author ID has better precision in terms of captured co-authorship than some previous approaches. When we want to use publication keywords, personal interests and conference data to construct an affiliation network, it is slightly more complicated. The publication keywords data we mined from EPrints can be formatted in various forms - each keyword can be separated by a space or different punctuation; same keywords may have different expressions (e.g. web2.0 and Web 2.0), a keyword may have extra annotation inside a pair of brackets or parentheses. The personal interests data is extracted from researcher online profiles, the data is stored in the *Resource Description Framework* format. The conference data mining require some basic intelligent inferring process, since the data we desire is the theme of the conference. For instance, in conference data -"The 7th IEEE International Conference on Advanced Learning Technologies (ICALT 2007)", the useful information we would like to have is "*Advanced Learning Technologies*" and "*ICALT*" (this is useful, as some authors only use abbreviations in their reference). Therefore, extracting useful information requires removing numbers, common stop words (e.g. "on"), general conference title (e.g. "IEEE conference") and punctuation and symbols. Sometimes, author may also include city names, country names and date in the "conference events" field, and may write name of number instead of number (e.g. seventeenth instead of 17th), all these special cases of data are removed and data is further split based on stop words and punctuation. After the data is processed mined data, it is then employed to construct an affiliation network. Here we adopt the tripartite model for folksonomies proposed by Mika [10] as a formal model of this network. Assume we have set of researchers $R = \{r_1 ...,r_n\}$, keywords, $K = \{k_1,...,k_m\}$ and publications $P = \{p_1,...,p_j\}$, and we can construct a tripartite graph $T \subseteq R \times K \times P$. Such a network is most naturally represented as a hypergraph with ternary edges, where the edge represents the fact that a given researcher associated a certain publication paper with a keyword. In particular, we define the hypergraph representing T as $H(T) = <V, E>$, where V is the set of vertices: $V = R \cup K \cup P$, and E is the set of edges $E = \{(r, k, p) \mid (r, k, p) \in T\}$. As described by Mika, the folksonomy tripartite graph can be reduced to a bipartite graph depending on the purpose of analysis. In our case, we are interested in the graph which is constructed by researcher and keywords. Therefore, by focusing on a single researcher, we could use the following definition to construct a researcher and keywords bipartite graph: RKr = $<R \cup K, E_{rk}>$, where $E_{rk} = \{(r, k) \mid (r, k, p) \in T\}$. This graph can be represented in matrix form, in which we denote the bipartite graph as B = $\{b_{ij}\}$, where $b_{ij} = 1$ if there is an edge connecting a researcher r_i and a keyword k_j, $b_{ij} = 0$ otherwise. We define a new matrix, S = BB^T, known as the co-affiliation matrix, which defines a social network that connects researchers based on shared keywords. The diagonal of this matrix contains the counts of how many

[2] http://www.imdb.com/

[3] http://www.oakland.edu/enp/

keywords a given researcher was affiliated with in the bipartite graph.

In order to measure if two researchers share similar research interests, we need to measure the distance between two sets of data which are collected to represent two researcher's interests. The similarity measuring algorithm created in this project will take into account the weight of each keyword as well. The algorithm works as below:

1) For each person in LSL, computes their own keyword list
2) Compare this with the each other person's keyword list by using the following formula: $L_{\text{matched keywords}} = \{ L_{\text{keywords1}} \} \cap \{L_{\text{keywords2}}\}$
3) For each keyword in $L_{\text{matched keywords}}$, compare its original paired weighting value (the frequency a keyword appeared in a researcher's publication) in $L_{\text{keywords1}}$ and $L_{\text{keywords2}}$
4) The smaller weighting value indicates the frequency of keywords matched ($f_{\text{keywords matched}}$)
5) The similarity value between these two person is: $V_{\text{similarity}} = \sum f_{\text{keywords matched}}$

For instance, if we use nested pairs (keywords, weight) in the list data structure to represent user A's research interests L_A = [(*semantic web*, 10), (*hypertext*, 4), (*folksonomy*, 1)] and user B's research interests as L_B = [(*e-learning*, 12), (*semantic web*, 8), (*hypertext*, 1)]. The two keywords matched are "*semantic web*", and "*hypertext*", where "*semantic web"* matched 8 times, and the "*hypertext*" matched only once. Therefore the similarity value is 9 between user A and B.

4. EVALUATION RESULT ANALYSIS

Based on the four types of metadata, social network graphs were constructed and analysed. In order to generate one "ideal" graph, we needed to estimate each of the data types' weight to blend all the data together. By the "ideal" graph, we mean that (1) a graph that represents all individual researchers' research focus; (2) a graph that captures the social relationships between researchers, and (3) by using the graph, users are able to retrieve the desired information. We proposed the following formula to compute the edge of the researcher social network graph:

$$E_R = \alpha K + \beta I + \gamma C + \delta CA \qquad (2)$$

E_R - *Edge of Researcher Social Network Graph;* K - *Publication Keywords;* I – *Interests;* C – *Conference;* CA - *Co-Authorship;*

By calibrating the coefficients α, β, γ, δ, we can generate different researcher social network graphs. An experiment was then designed to evaluate social network graphs and statistically estimates the coefficients α, β, γ, δ,

The experiments were constructed with a questionnaire, a graph comparison exercise and an informal interview. Based on the different purposes of evaluation, the experiment was divided into four sections. In the first section, a single participant's EPrints publication keywords and ECS info page personal interest data are collected and meshed together in advance. During the experiments, we asked the participants to select up to 10 keywords which represent their research focus. By comparing the participants' selected keywords and their original data, we were able to statistically estimate the weighting of the coefficients between EPrints keywords and interests stated on their personal profile. On average, there are about 7.8 out of 71.6 keywords and 3.2 out of 7.5 personal interests selected by a single participant. If a word belongs to both keywords and personal interest data, it counted as 1 from both. The keywords selected are about 2.5 times greater than personal interests selected. This is interesting because the personal interests data are explicit statement about their research interests, whereas the keywords extracted from research papers indicates a scholar's actual research focus. If we examine the keywords and interests further, in terms of quantity, EPrints keyword data is about ten times greater than interests. In terms of expressiveness, the words that the researchers used to describe their research interests tend to be more abstract (*e.g. Hypertext*), whereas the keywords obtained from EPrints are more elaborate (*e.g. adaptive hypertext*). In terms of validation, EPrints keywords are constantly enhanced as more new publications are added to the archive and also time stamped with the paper publication dates, which therefore can be easily filtered and selected, whereas interests are not updated within time. Furthermore, during the experiments, synonyms were considered as different keywords, otherwise keywords would score even higher rates. However, the personal interests do capture certain personal attributes which publication keywords do not. For instance, a computer scientist may have never published any research paper about astronautics but still states his or her interest in the topic. From the perspective of building a cross field collaboration of scientific social network, this data could be valuable. If we compare the coverage of data, only 11% of the keywords data are selected from the publication keywords collection, but 42.7% of the interests were chosen from personal interest collection. From this point of view, the interests may not be abundant, but it has a higher precision percentage than keywords. In the second section, four graphs were shown to the participants (without stating what data the graphs were based on). The participants were required to pay attention to the people positioned close to them, to circle people who share similar research interests, and to cross out people who do not. This was designed to estimate the accuracy of each graph. The results are shown in Table 4.1, where the co-authorship graph has a high accuracy rate of 71.5% and is considered the highest precision graph amongst the others. These results are consistent with the experiments described in [9]. One factor that may affect the results is that the participants tend to select people more familiar to them. There were more selected people on the co-authorship graph in comparison with other graphs, since the participants should know the majority of people who published a paper with them. The

10.5% unknown (the researchers on the graph are not circled or crossed), therefore, tend to be the people they do not want to judge, rather than indeed they are unknown.

Table 4.1 – 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%

Connected People Circled - Participants agree that people on the graph share similar research interests

Connected People Crossed - Participants do not agree that people on the graph share similar research interests

Unknown - People who are either not circled or crossed but still have a connection with participants on the graph

Comparing with the higher unknown rate of keywords, interest and conference data, those percentages are more likely to represent people whose research interests which they really do not know. Therefore, we cannot conclude that keywords, interests and conference data indicate lower precision results. Section three required the participants to rate on a scale of 1 to 10 the representativeness of the graph and the importance of data in terms of the scholar's research interests. The result was then normalised by ranking order. The network constructed based on conference data was ranked as number one and interests were considered the worst in terms of representativeness. Keywords were considered as the most important data, whereas the conference data was ranked as least important. The majority consider the keywords have high precision to capture research interests as it reveals what research area one is working on. Conference data is more disputable, some participants consider that "you may go to a conference that is just relating to your research area." The positive side assume that - "Conferences are expensive in time and money to go to. If you attend a conference, it represents your genuine interests, unless it is a beautiful place that somebody wants to go for a holiday." Where some neutral voices state that conference data would have a higher value if the conference data is consistent (a researcher attends a conference regularly). Based on ranking results, the estimation of the coefficient is ordered as table 4.2.

Theoretically, the conference data extracted from EPrints is not elaborate and have a high degree of noise, making it less than ideal as an indicator of researcher interests. However, the graph was considered the best in terms of representiveness of a researcher's interests. The simple explanation is that, with limited size of publication data, the social relations captured by using conference data extracted from references contains the relations which are captured by using co-authorship approach (only applicable for paper published in the conference). As for a single piece of data, we extracted the conference data for all co-authors to be used to match with other researchers, however, this also established the similarity relationship between co-authors themselves. Therefore, the conference graph is more like a partial co-author graph with an extra amount of matching relations. As the 7682 publications used in this project represent a relatively small dataset, it is not sufficient to prove that the conference graph will always be better than the co-author graph. If we deployed such a social network on a global scale, the results would be different - as thousands of researchers go to the same conference every year, the co-authorship attributes weight will become less significant by comparison.

Table 4.2 – Rating of importance of data in terms of how it represent a scholar's research interests

Evidence	**Estimation of weighting coefficient**
Ranking of Representative of graph	$\gamma > \alpha > \delta > \beta$
Ranking of Importance of Data	$\alpha > \delta > \beta > \gamma$

α - Coefficient of Eprints Keywords; β - Coefficient of ECS Info Interests; γ – Coefficient of conference Attended; δ – Coefficient of co-Authorship

If we further examine the information extracted from EPrints publication keywords, personal interests, and conference data, all this information is used to represent a single scholar's research interests. It could also be understood that this data is employed to describe one's personal attributes, where information provided by co-authorship is the relationship between researchers, which is summarised as one type of their social relations. When we construct a social network based on personal attributes, only vertexes are added and edges are added by executing the algorithm to compute the union set of two different personal attributes. In contrast, a social relation already represents a single edge on the graph; the graph is constructed by the accumulation of edges. Based on the difference of the two approaches, we propose a dynamic social network data model:

$$E_{social\ network} = X \times E_{personal\ attributes} + Y \times E_{social\ relations} \quad (3)$$

E – Edge of Graph

By calibrating the weighting on each set of data, we could generate a researchers' social network for different purposes. Based on the result in table 4.2, using keywords is considered better than co-authorship in both graph representative and the importance of the data. Therefore, we believe that setting a higher weight (e.g. X ≥ Y) on the personal attributes dataset than in social relations will produce a better scientific collaboration social network which captures the similarity in research interests. Figure 4.1 shows a graph based on blending keywords, interests, conference and co-authorship; the corresponding weight are $\alpha = 0.32$, $\beta = 0.13$, $\gamma = 0.35$ and $\delta = 0.2$, where $\gamma > \alpha > \delta > \beta$.

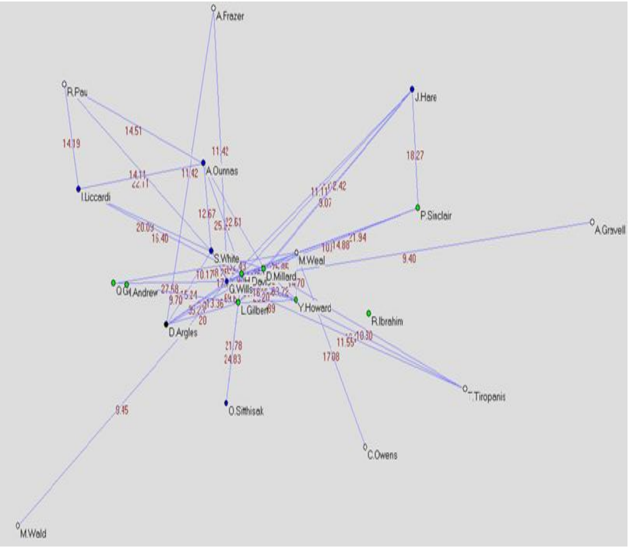

Figure 4.1 Researcher Social Network Graph Based on Blended Data (Edge threshold > 8)

5. CONCLUSION

The results from the researcher social network evaluation experiments suggested estimated weights for the different types of data in order to construct an ideal social network graph. Publication keywords data were considered about 2.5 times more important to represent one's research interests than explicitly stated personal interests' data. Both the ranking of data importance and the ranking of data representativeness, as visualized in the graphs, indicated that the publication keywords data should have more weight than the co-authorship data, where co-authorship data should be heavier than personal interests' data. It is worth emphasising that to structure a social graph by only using keywords 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 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. Different results are expected when a group of researchers of more diverse research interests are analysed, such as the members of a university or of a national research council. Finally, we propose a dynamic researcher social network data model ($E_{social\ network} = X \times E_{personal\ attributes} + Y \times E_{social\ relations}$) which categorizes data into two groups: personal attributes (e.g. publication keywords) and social relations (e.g. co-authorship). By calibrating the weighting on each set of data, we could generate a researchers' social network for different purposes. Based on this project, a new experiment could be designed to evaluate the representativeness of a hybrid personal attributes graph (keywords, interests and conference) and the similar experiments would be carried further between two different research groups or two different research fields (e.g. computer science and social science), for the purpose of exploiting the value of the network for cross field collaboration. Data mining for the network could be scaled up to a larger area to evaluate the potential problem of a researcher social network on a global scale, for instance to harvest the Association for Computing Machinery (ACM) publication metadata to construct a network. The empirical results of this research could be commercially used to implement global scale researchers' social network applications and publication recommendation systems.

6. REFERENCES

[1] T. M. Harrison, and T. D. Stephen, "The electronic journal as the heart of an online scholarly community - Networked Scholarly Publishing," *Library Trends*, 1995.

[2] M. E. J. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences*, pp. 404-409, 2001.

[3] T. J. Fararo, and M. Sunshine, "A Study of a Biased Friendship Network," *Syracuse University Press*, 1964.

[4] "The Oracle of Bacon at Virginia," 09/24, 2008; http://www.cs.virginia.edu/oracle/.

[5] M. E. J. Newman, "Scientific collaboration networks: I. Network construction and fundamental results," *Physical Review*, vol. E 64, 2001.

[6] J. W. Grossman, "The Evolution of the Mathematical Research Collaboration Graph," *Congressus Numerantium*, vol. 158, pp. 202-212, 2002.

[7] A. L. Barabási, and R. E. Crandall, "Linked: The New Science of Networks," *American Journal of Physics*, no. 71, pp. 409, 2003.

[8] J. W. Grossman, and P. D. F. Ion, "On a Portion of the Well-Known Collaboration Graph," *Congressus Numerantium*, vol. 108, pp. 129-131, 1995.

[9] Y. Matsuo, H. Tomobe, K. Hasida and M Ishizuka, "Mining Social Network of Conference Participants from the Web " *Proceedings of the International Conference on Web Intelligence*, 2003.

[10] P. Mika, "Ontologies are us: A unified model of social networks and semantics," *Journal of Web Semantics*, vol. 5, pp. 5-15, 2005.