
Analysis of network structures to support scientific collaboration

Proposal submitted by F. Heylighen to the Belgian Fund for Scientific Research
Requested funding: 1 4-year PhD scholarship + 7700 euro per year for costs and equipment.

PREVIOUS RESEARCH

The "Evolution, Complexity and Cognition" group within the Center Leo Apostel has over ten years of experience in using network models to analyse socially distributed knowledge, and thus to support collective information retrieval and problem-solving [Heylighen, 1999]. This approach has been applied mostly to the World-Wide Web and electronic document collections, using the ways users traverse hyperlinks between webpages to find which pages are related to each other [Bollen & Heylighen, 1996; Heylighen & Bollen, 2002]. A more recent application is societal-scale collective decision-making via the network-based delegation of voting powers [Rodriguez & Steinbock, 2004].

AIM

We are entering an era where society is being exposed to the idea that patterns of human relationships can be understood as networks [Barabasi, 2002]. As computer networks provide the technical infrastructure for the explicit representation of social relations in a shared collective space, we are seeing the rise of a wholly new application of social network analysis methods: societies analyzing themselves on a large scale, affecting a kind of collective self-reflection. The key point is this: when a society gains access to a representation of its own network structure, albeit imperfect and approximate, this knowledge can act as a causal agent in the evolution of that structure. In other words, social networks are maps of social territory which are useful for individuals navigating that territory. We navigate social space whenever we search for individuals or groups meeting some criteria—experts, collaborators, friends—with the intention of forming new network ties.

For the purposes of this project, we will apply algorithms to the social networks created by the research community. When two scientists collaborate on a paper, the publication they produce creates a tie between them in the research communities co-authorship network. Co-authorship networks have been studied extensively for their network properties [Newman 2001a, 2001b, 2003; Watts, 2002; White, et.al., 2004], but little work has been done to exploit this information for a practical application. These implicit networks are made explicit thanks to the large online publication repositories such as CiteSeer, arXiv, and e-print servers in general. These repositories are increasing in popularity as institutions are beginning to see the cost benefits of these

media. For example arXiv's repository is growing at a rate of 30,000 manuscripts a year [Garner, et.al., 2001]. This project wishes to take advantage of this infrastructure to support the self-reflective development of the scientific community. Through meta-level analysis of the group's relationships it is possible to better organize the flow of information between scientists and foster potentially synergistic relationships.

OBJECTIVE 1: RECOMMENDING PUBLICATIONS

Publication dissemination is an essential part of research. Papers require a public forum if they are to be used by the community. Journals and conference proceedings have for a long time been the main method by which the community gets access to research findings. Unfortunately, because of restricted circulation and high fees, this information is provided only to the few subscribing to, or attending, these various forums. This situation has been remedied by the use of wholly or partly free online publication repositories. Modern advances in online publishing have received much attention and application [Garner, et.al., 2001] [Menon & Raymond, 2002] [Bollen & Luce, 2002]. Online paper repositories provide the general public with search mechanisms to find papers in various disciplines. Furthermore, with the very recent release of Google Scholar and its ability to aggregate the information from multiple repositories, ordered by impact to the field, locating papers that are useful to one's own research is becoming increasingly easier.

However, the downside of this is an explosion in the amount of available information. This requires effective tools to filter information, so that only the documents most relevant for a particular researcher's interests are retrieved. Methods that utilize e-prints technology allow individuals to subscribe to particular manuscript keywords in order to allow the server to perform pattern matching then delivery [Garner, et.al., 2001]. But keywords are in general not sufficient, as they still produce plenty of documents of low relevance, while ignoring important work because it uses a different terminology [Heylighen & Bollen, 2002]. By exploiting the social-network structure of the community and the implicit interests it expresses, we wish to design a system that recommends potentially useful documents based on the collective preferences of the community [Heylighen, 1999].

OBJECTIVE 2: PROMOTING COLLABORATION

Our second objective is to create collaborative links within the community. Instead of just routing information to individuals, we wish to recommend individuals to other individuals as potential collaborators for projects that have yet to be started. Both objectives intend to stimulate a more interconnected and well-informed scientific community by uncovering associations implicit in the existing network of authors and documents.

It has been argued that institutional boundaries are a major impediment to the collaboration of scientists that may in fact have a fruitful relationship [Becher, 1989]. These boundaries come in the form of various coarse-grain academic disciplines as well as institutional boundaries such as those between academia and the commercial sector. Thus far, the most common method to cross these boundaries is publication. After publication manuscripts from one domain or institute become accessible to everyone, regardless of research domains. Still, as noted in the previous objective, this information is not always able to locate those individuals which may be interested in the work since it requires an active search by potential readers. Even with advances in information dissemination, the methods to create collaboration prior to publication are still relatively unaddressed [Laudel & Gläser, 1998].

We aim to use the network structure of the scientific community as a tool to locate and recognize researchers working on related topics, independently of institutional or disciplinary boundaries. Thus, individuals using our system would receive a list of other researchers they might want to contact for possible collaboration. This can go from the simple exchange of preprints or email, to the location of peers to review a paper, consultation about technical problems, and collaboration on concrete projects. On the larger scale, this would be useful for creating Networks of Excellence and international consortia for submitting proposals to European Union funding programs. In this way, communities can be formed around particular topics, united by their interest.

STARTING ASSUMPTIONS

We conceive the scientific world as a huge network consisting of two types of components: individuals (researchers) and documents (papers, books, web pages...). These are linked to each other, as each document has one or more individual authors. Documents are moreover linked to other documents, as one document typically cites others through its list of references. These relatively sparse direct references are complemented by the indirect connections that can be derived from them: co-citation (two documents are linked when they are referred to by the same third document) and bibliographic coupling (two documents are coupled when they both refer to the same third document). Authors are also indirectly linked to each other via the documents they have written (co-authorship, or one author citing another one). These links represent a form of mutual trust, knowledge or respect, thus defining a social network [White, et.al., 2004]. By suitably combining the different direct and indirect connections [Heylighen & Bollen, 2002], we can derive one network of authors and documents, where every component has multiple, weighted connections to other components that express their direct and indirect associations. The next step is to exploit these connections to recommend individuals or documents to other individuals.

METHODOLOGY

To analyse such a network, we can use different recurrent algorithms, which repeatedly check connections and connections following connections, thus producing an overall measure of the degree of direct and indirect relatedness between two nodes in the network. Perhaps the best known of these techniques is spreading activation, whereby the nodes of interest receive a certain amount of activation, and that activation is propagated in parallel along the connections, proportionally to the weight of the connections [Heylighen & Bollen, 2002]. This activation typically diffuses through the network, decaying the further it propagates. However, different activation patterns can converge on the same node, producing a "constructive interference" so that activation accumulates in this point or region, suggesting that it is strongly associated with the initially activated node(s). This method can help us to uncover clusters of documents or researchers that form a relatively closed community or cliques, referring mostly to each other but rarely to outsiders.

A related method uses "particle" diffusion over a network. The difference is that activation can decay continuously, whereas particles are discrete and are either transferred to another node, or not. This method has been applied in our group to search social networks so as to delegate voting powers [Rodriguez & Steinbock, 2004]. With this method, particles may encapsulate URL references to papers. The more particles an author in the network accumulates that maintain the same URL, the higher the probability that that author will indeed find the paper of interest.

A very recent paper proposes a number of related new mathematical methods to determine the similarity or relatedness between nodes in a network [Fouss, et.al., 2004]. The main idea here is that the relatedness increases when the connecting paths between nodes become shorter and/or more numerous. For example, although two researchers may never have directly collaborated, or not even shared collaborators, their collaborators may all belong to the same community so that there are many indirect collaboration connections. This method, which is related to the Katz [1953] measure in social network analysis, has proven its reliability in tests using data for recommending movies to moviegoers [Fouss, et.al., 2004].

We will apply these different methods to a network derived from available on-line authorship and citation data, restricted to a suitably chosen, not too large domain (e.g. condensed matter physics, cf. [Newman, 2001]). We will select a number of authors at random from this network, and then generate for each one of them a list of recommended papers and potential collaborators. We will then send this list to the authors by email, together with a survey asking them to indicate how useful they consider these recommendations to be. This will allow us to see which methods produce the best results.

In order to test these results against current keyword-based distribution methods, in a second stage, for each author we will also produce a list of 3-4 characteristic keywords, which we either ask them to suggest to us before they get the recommendation, or distill from the keywords they use in their publications. These keywords will be used to find papers using existing search engines (e.g. Google Scholar),

and suggest possible collaborators from the authors of these papers. These keyword-based recommendations will then be mixed with our network-based ones, and sent to the authors without them knowing which is which. The scores of the different recommended items on the survey will then tell us to what degree or in what respect the network-based method is better or worse than the keyword based-one. Statistical T-Tests will be used to determine the significance of these results.

RESEARCH STAGES

2006: in the first year, we will explore different on-line repositories for publication data to generate suitable author-document networks. By performing preliminary tests on these networks, we determine the optimal size and coverage to make our further experiments both realistic and not overly complex.

2007: we apply the different algorithms to generate indirect relations, and compare the results by means of a survey of authors receiving the resulting recommendations.

2008: we select the best performing network-based recommendation algorithm, compare its results with keyword-based recommendations by holding a second survey, and analyse the differences.

2009: we integrate the different findings into a general model which is published under the form of international publications and a PhD thesis. The network analysis software that we developed is made freely available on the web for other researchers to use, both to develop further and to directly get recommendations for themselves.

CONCLUSION

Combining social-networks and recurrent network search algorithms is the novel contribution that this project proposes. This technique provides a passive means of information distribution that is as of yet unseen. If this method proves to be successful at routing information within author-document networks, then generalizations to other social-network applications can be envisaged. Together these two techniques of information dissemination and community building will lead to deliverables in the form of online tools that utilize online publication repositories to generate additional recommendations for any individual interested in one or more publications or scientists. It is interesting to note that scientists need not do any effort to enter data into the system since this information is already available through the online publication repositories. This is a common problem with much social software that requires a critical mass of users willing to participate before the tool becomes useful. In fact, scientists can be automatically recommended publications and potential collaborations without even being aware of the system's existence. It will be up to us as developers to ensure that privacy issues are not violated when using such solicitation techniques.

REFERENCES

- Barabasi, A.,L., "Linked: The New Science of Networks", Perseus Publishing, 2002.
- Becher, T., "Academic Tribes and Territories. Intellectual enquiry and the cultures of disciplines", Milton Keynes: Open University Press, 1989.
- Bollen, J., Luce R., "Evaluation of digital library impact and user communities by analysis of usage patterns", D-Lib Magazine, vol. 8:6, 2002.
- Bollen J. & Heylighen F. (1996) "Algorithms for the Self-organisation of Distributed, Multi-user Networks", in: *Cybernetics and Systems '96* R. Trappl (ed.), (Austrian Society for Cybernetics), p. 911-916.
- Fouss, F., Pirootte, A., Renders, J., Saerens, M., "A novel way of computing dissimilarities between nodes of a graph, with application to collaborative filtering and subspace projection of the graph nodes". European Conference on Machine Learning Proceedings, ECML, 2004.
- Garner, J., Horwood, L., Sullivan, S., "The place of eprints in scholarly information delivery", Online Information Review, vol 25:4, pp. 250-256, 2001.
- Heylighen F. & Bollen J. (2002): "Hebbian Algorithms for a Digital Library Recommendation System", in *Proc. 2002 Int. Conference on Parallel Processing Workshops* (IEEE Computer Society Press)
- Heylighen F. (1999): "Collective Intelligence and its Implementation on the Web", *Computational and Mathematical Organization Theory* 5(3), p. 253-280.
- Katz, L. A new status index derived from sociometric analysis. *Psychometrika* 18(1), 39-43, 1953.
- Laudel, G., Gläser, J., "What are the Institutional Boundaries and how can they be Overcome? Germany's Collaborative Research Centres as Boundary-Spanning Networks", *Cultures of Science and Technology – Europe and the Global Context*, EASST, 1998.
- McPherson, M., Smith-Lovin, L., and Cook, J., "Birds of a Feather: Homophily in Social Networks", *Annual Review of Sociology*, 27415-44, 2001.
- Menon, G.M., Raymond, F.B., "Electronic Publishing: An avenue for the dissemination of knowledge", *Electronic Journal of Social Work*, vol. 1:1, 2002.
- Newman, M., "Scientific collaboration networks. I. Network constructions and fundamental results", *Physical Review*, vol. 64, 2001a.
- Newman, M., "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality", *Physical Review*, vol. 64, 2001b.
- Newman, M., "Ego-centered networks and the ripple effect", *Social Networks*, vol. 25, pp. 83-95, 2003.
- Rodriguez, M., Steinbock, D., "A Social Network for Societal-Scale Decision-Making Systems", *North American Association for Computational Social and Organizational Science Conference Proceedings*, 2004.
- Watts, D., Dodds, P., Newman, M., "Identity and search in social networks", *Science*, vol. 296, pp. 1302-1305, 2002.
- Watts, D., "Small Worlds: The Dynamics of Networks between Order and Randomness", Princeton University Press, 2003.
- White, H., Welleman, B., Nazer, N., "Does citation reflect social structure? longitudinal evidence from the "Globenet" interdisciplinary research group", *Journal of the American Society for Information Science and Technology*, vol 55:2, pp. 111-126, 2004.

WORLD WIDE WEB REFERENCES

- arXiv Online Publication Repository: <http://www.arxiv.org/>
- Citeseer Online Publication Repository: <http://citeseer.ist.psu.edu/>
- Eprints Open Source Project: <http://www.eprints.org/>
- Google Scholar: <http://scholar.google.com/>