

# Towards Cross-Community Effects in Scientific Communities\*

[work in progress]

Marcel Karnstedt and Conor Hayes  
Digital Enterprise Research Institute (DERI)  
National University of Ireland, Galway, Ireland  
first.last@deri.org

## Abstract

Community effects on the behaviour of individuals, the community itself and other communities can be observed in a wide range of applications. This is true in scientific research, where communities of researchers have increasingly to justify their impact and progress to funding agencies. Previous work has tried to explain these phenomena by analysing co-citation graphs with methods from social network analysis and graph mining. More recent approaches have supplemented this with techniques from textual clustering. However, there is still a great potential for increasing the quality and accuracy of this analysis, especially in the context of cross-community effects. In this work, we present existing approaches and discuss their strengths and weaknesses. Based on this, we choose two closely related communities and propose novel ideas to detect and explain cross-community effects with a special focus on their characteristics in a given timeline. The outcome is a roadmap for advanced analysis of cross-community effects, which promises valuable insights for all areas of scientific research.

## 1 Introduction

Community structures can be found in a wide variety of applications. Analysing these structures provided very interesting insights into the internals and functioning of social networks since first works in this field appeared. This began with Milgram's famous experiment on the "six degrees of separation" [Milgram, 1967] and can be found in the whole research area on social network analysis [Granovetter, 1973].

In current days, this topic experiences amplified attraction again. This is due to the wide variety and great potential of social applications in the Web, such as Facebook<sup>1</sup> and Wikipedia<sup>2</sup>. Researchers as well as economists expect valuable outcomes for optimising Web technologies and increasing revenue from the detailed analysis of community structures and their effects. An evidence for the popularity of methods from social sciences is the existence of the "six degrees of Kevin Bacon"<sup>3</sup>, in relation to Milgram's orig-

\*This material is based upon works jointly supported by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2) and under Grant No. 08/SRC/I1407 (Cliques: Graph & Network Analysis Cluster)

<sup>1</sup><http://www.facebook.com>

<sup>2</sup><http://www.wikipedia.org>

<sup>3</sup><http://www.thekevinbacongame.com>

inal work. A main problem of applying these techniques is their usually restricted scalability. Network structures found in the current Web tend to be by far too large for directly applying these methods, which are developed for small graphs around single individuals.

One specific field in this area, which is especially interesting for computer researchers, is the analysis of scientific communities. The practice of citing other authors' works is particularly typical for computer scientists. The detection and explanation of community effects helps to justify progress and gather funding as well as to identify trends and evolving fields over time. It can guide funding agencies and tenure committees to make more informed decisions. It has been shown that citation analysis is a very promising approach to detect correlations and interdependencies between different researchers and fields of research.

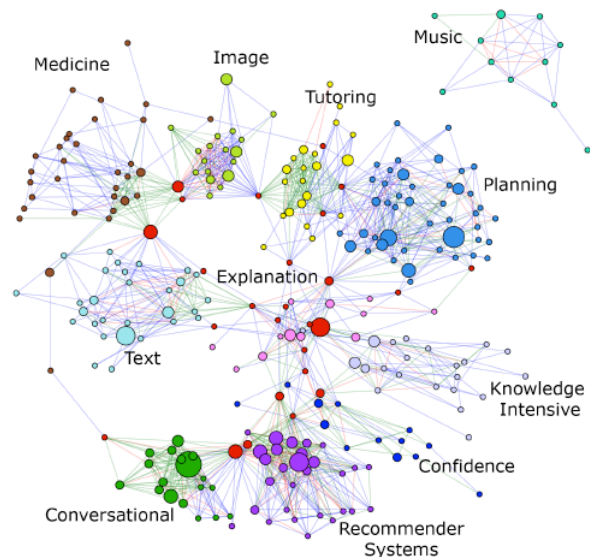


Figure 1: A co-citation graph for social network analysis [Greene *et al.*, 2009]

So far, most of these works focused on specific communities and the different sub-communities, i.e., different fields in one general area of computer research. But, the analysis of effects between different broader communities promises to reveal more and different insights. This can help to leverage inter-community communication and collaboration as well as to increase the impact of research. To approach this novel view on community analysis, we choose two closely related communities as a starting point. Our choice falls on the Semantic Web community, as a rather young but dramatically evolving one, and the Information Retrieval community, as a profiled and closely related community. Later, we plan to extend the analysis

to other related communities, such as that from Database research. The goal is to develop techniques for analysing cross-community effects for an arbitrary number of communities. We expect valuable insights not only for the chosen fields of research, but also for scientific research in general. The outcomes of this work will have great potential for understanding, directing and optimising the effects and interdependencies between the identified groups of scientific research.

The general approach starts with collecting a set of seed papers. For these works, citations have to be extracted. Based on this, a graph of citation relations can be build, where the vertices represent works or authors and the edges between nodes represent relations between them based on citations. Figure 1 shows an example of such a graph, taken from [Greene *et al.*, 2009]. The graph shows the community structure in the field of Case-Based Reasoning (CBR) in the state of the year 2008. The figure shows the structure of one community, but also indicates how the structure between communities could look like. On top of such a graph, methods from social network analysis and graph mining will help to understand the specific cross-community effects. In Section 2, we give brief insights into the phenomena we expect. Finally, we provide a roadmap for following works, whereby we focus on the three main aspects of data gathering (Section 3.1), graph construction (Section 3.2) and actual analysis (Section 3.3). Section 4 concludes the paper.

## 2 Expected Phenomena

There is a wide range of phenomena that we expect to find with the methods proposed in this work. In this section, we briefly describe two selected ones, namely the paradigm shift [Kuhn, 1996] and paradigm merge. These are two effects that play a specific role especially for the analysis of scientific communities and the cross-community relations between them.

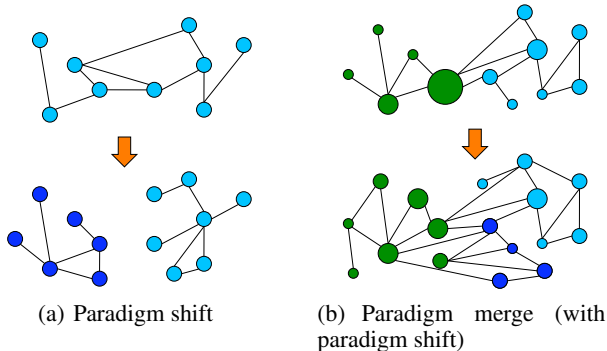


Figure 2: Paradigm shift and paradigm merge as possible phenomena

Figure 2(a) illustrates what can be called a paradigm shift in scientific communities. The upper part shows citation relations between different authors or works that might be found at a specific point in time. Analysing the development of this graph over time might reveal that a sub-community somehow detaches from its original community. This means, authors from both communities do not cite each other any more, with ongoing time the sub-community seems to “speak a different language” that is not understood by the remaining community any more. Such a phenomenon was first described by Kuhn and called a paradigm shift [Kuhn, 1996]. Clearly, to detect such an effect, the citation structure has to be analysed over time, by

looking at the corresponding graphs from different points in time.

In Figure 2(b) we show the opposite of this, which we call a paradigm merge. Such an effect can be expected particularly when analysing two or more originally separated communities. Over time, the communities approach each other, represented by more and more citings between them. This can lead to closely related communities or even to a merge into one larger community. For some communities, we even expect a combination of paradigm shift and merge. This means, from one large community only a sub-community approaches the originally separated one – whereby in parallel detaching from its original field of research. We indicate this in the figure by the different shades of the nodes.

Especially for new and rapidly evolving communities like the field of Semantic Web research we expect this to be observed in combination with another effect. In its beginning, the only work that is “visible” to other communities might be a very fundamental and ground-breaking contribution by one of its founders. For the Semantic Web, one can think of Tim Berners-Lee famous work [Berners-Lee *et al.*, 2001], which is seen as the initial work founding that community. We indicate such a visibility by using differently sized nodes in the figure. Over time, more works “appear on the horizon”, the coastline of the community becomes visible and more islands are cited. This results in a shift of visibility between the actors of the evolving community. Analysing such effects can be done by looking at the citation graph accumulated over time (i.e., the graph can only grow) or by analysing graphs from different points in time.

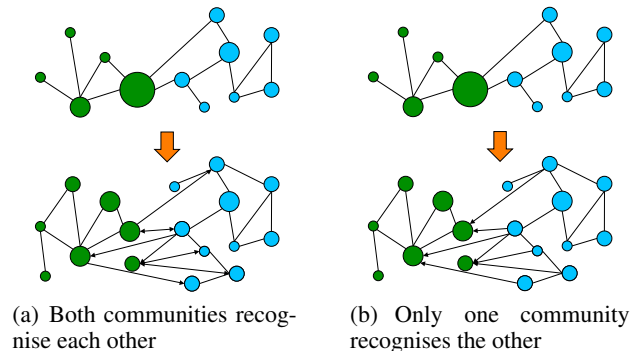


Figure 3: Communities may both recognise each other or one can reveal a “non-social” behaviour

Regarding the phenomena of paradigm shift and merge, there are also other aspects we plan to take into account. One of the communities might show a “non-social” behaviour, simply neglecting the existence and development of new communities. In this case, we expect only one community to cite the other. In contrast, a healthy development would be observed if both communities increasingly cite each other over time. Figure 3 illustrates that difference by using directed edges that indicate the direction of citations. This leads to other important questions, such as what are (un)healthy communities and how to detect that.

Note that the effects are illustrated here in a rather dramatic manner. We expect these phenomena to usually occur alleviated. This means, rather than only one community citing only the other (Figure 3(b)), one community might cite the other in a much more intensive manner. One can see this as this community having more tentacles in other communities. In contrast, certain fields might tend to cite only the “tall” figures visible, even if the community that these figures belong to matures.

### 3 Roadmap

As outlined before, we focus on citation analysis as the tool of choice for detecting and understanding cross-community effects. It has been shown that this is a reasonable and well-functioning approach for this task. In this section, we provide a general roadmap that defines the different tasks we have to fulfill to achieve our goals. We first discuss how to gather and prepare the citation data needed. Afterwards, we present ways for building social network graphs on top of this data and finally get over to actual methods we aim to apply on the so built networks.

#### 3.1 Data Gathering

First works in the area of citation analysis [White and Griffith, 1980; Gmür, 2003] had to rely on specialised citation databases like the *Social Sciences Citation Index* (SSCI) for gathering the required data. Such databases reveal several disadvantages, such as pruned author lists and only a selection of papers. Luckily, today there are much more citation sources available. Sites like DBLP<sup>4</sup> and Springer<sup>5</sup> are well suited to select a set of seed papers. They also provide ways for selecting high-impact journals and conferences that specifically relate to the chosen communities. Based on this, sites like Google Scholar<sup>6</sup> and CiteULike<sup>7</sup> can be used to extract according citation data, without the need for parsing the chosen papers. The social aspects of, for instance, CiteULike that supports tagging and grouping of works one prefers, further help to identify topics and fields of research. Usually, the raw input data has to be cleaned (different usage of author names and paper titles) and probably pruned to the most significant (most cited) works. We will evaluate if this is also suited for analysing cross-community effects, where we may also be interested in rather small islands of community landscapes.

With most existing approaches, the results of the citation analysis have to be inspected manually in order to deduce meaningful results. [He and Hui, 2002] is one of the first works aiming at automising the whole process. However, the labeling of sub-communities has still to be done on the basis of human inspection. In order to also automate this process, we plan to use a tagging approach. The generated tags can be used to identify and name the found areas of research. To achieve this, one approach is to use already provided keywords and methods from Natural Language Processing. However, we expect this to be too inaccurate and not absolutely satisfying. A second approach that we plan to use is to use input from the communities themselves. For this, a kind of game (similar to the ESP game<sup>8</sup>, also adopted by the Google Image Labeler<sup>9</sup>) could be supported. Another idea is to provide an interface for tagging own works (“eating their own dogfood” – a phrase popular in the Semantic Web community). First experiences will show whether these ways for generating tags are sufficient or not.

#### 3.2 Building Citation Graphs

Based on the raw citation data, we will have to build a graph of citation relations. Co-citation analysis has proved to be most suitable for this task. A co-citation between two

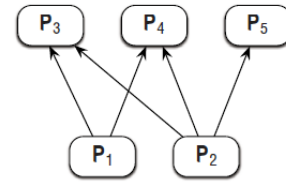


Figure 4: Principle of co-citations [Greene *et al.*, 2008]

works (i.e., an edge between two works or authors) is existing if both papers are cited in the same third work. The assumption is that if a co-citation link exists, the works can be regarded as very closely related in the same field of research. Figure 4 illustrates this principle. In this figure,  $P_3$  and  $P_4$  have a stronger co-citation relation than  $P_3$  and  $P_5$  and  $P_4$  and  $P_5$ . If  $P_2$  would not cite  $P_3$ ,  $P_3$  and  $P_5$  would have no co-citation link.

There are several different approaches for co-citation analysis. They mainly differ in:

- use a document-based or author-based approach (the nodes in the graph)
- whether to use absolute co-citation counts or relative values like Pearson’s correlation coefficient as in [He and Hui, 2002]
- macro vs. micro approach

Document-based analysis provides a more detailed view, but might assign an author to several (sub-)communities. However, we expect the document-based approach as more suitable for our needs, as it focuses on topics rather than geography as the author-based approach does. Usually, relative values can be expected to provide a more realistic view due to their normalising effect. The macro approach focuses on the overall structure of disciplines, whereas the micro approach tries to explain the structure and historical development of single disciplines. See [Gmür, 2003] for a good overview of the different approaches. [Gmür, 2003] also compares different approaches for clustering the citation data. We will evaluate the different methods with respect to their suitability to the special aims and expected phenomena in cross-community analysis.

However, applying only co-citation analysis is not sufficient. To handle aspects mentioned in Section 2, such as only one tall visible figure or direction of citations, we will have to apply pure citation analysis as well. A crucial task is to identify a good mixture of both and how to map the different techniques of co-citation analysis to the case of pure citation analysis. To the best of our knowledge, up to now no work aimed at combining both approaches.

#### 3.3 Analysing Citation Graphs

If we once built the citation graph, methods from social network analysis and graph mining seem to be most promising to analyse it. Maybe one of the most interesting methods is to apply different centrality measures. Eigenvector centrality and degree centrality have been shown to be especially suited [Greene *et al.*, 2008]. In general, we aim to identify bridges, hubs and further nodes of central importance. This might go along with a look on outer-world effects, as there may be bridges that actually belong to a third community. No existing work focuses specifically on effects between predefined communities and the influences of outer-world instances.

To reveal the inter-relationships among authors or works, three different approaches for multivariate analysis have been shown to be especially suited [He and Hui, 2002]:

<sup>4</sup><http://www.informatik.uni-trier.de/ley/db/index.html>

<sup>5</sup><http://www.springer.com>

<sup>6</sup><http://scholar.google.com>

<sup>7</sup><http://www.citeulike.org>

<sup>8</sup><http://www.espgame.org/gwap>

<sup>9</sup><http://images.google.com/imagelabeler>

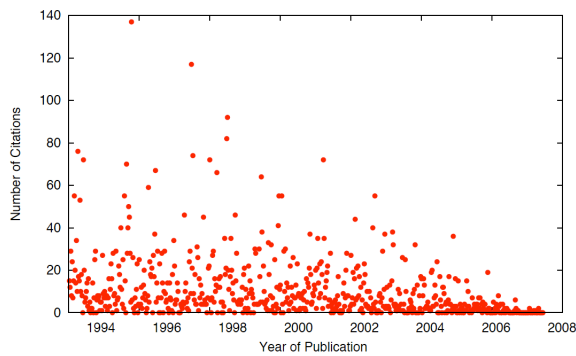


Figure 5: Citation counts over time [Greene *et al.*, 2009]

cluster analysis, multidimensional scaling (MDS) and factor analysis. Cluster analysis builds tree-like cluster representations, either following a top-down or a bottom-up approach. The MDS approach is used to build a map of authors or works, where heavily co-cited authors appear close to each other. MDS is especially suited for visually analysing the communities. More recent approaches for visualising social networks [Henry and Fekete, 2007; Gaudin and Quigley, 2008] will also be applied and evaluated. Factor analysis aims at defining a set of factors that authors contribute to, in their number much smaller than the number of nodes in the graph. Factor analysis has the advantage that it is able to assign authors to more than one factor, i.e., to more than one (sub-)community. [Greene *et al.*, 2008] proposes an interesting combination of hierarchical clustering and factor analysis, called *soft hierarchical clustering*. A detailed evaluation of the different approaches is part of our future work.

It is essential to use citation and co-citation counts to analyse (i) the structures in one community as well as (ii) the relations between different communities. Otherwise, we will not be able to identify things like a paradigm shift in combination with a paradigm merge as illustrated in Figure 2(b). For instance, looking at Figure 1, one can expect the *Explanation* sub-community as most visible to other communities outside CBR. We are interested in such sub-community effects that come along with cross-community effects.

As mentioned before, we are especially aiming at the analysis of time effects, i.e., the development of citation graphs over time. This involves analysing these graphs accumulated over time as well as in different points of time. It refers to the differences in the link structure as well as to changing positions of authors on an author map [White and Griffith, 1980]. Other time effects must not be ignored as well. For instance, it is natural that older papers are cited more often with time passing. This might be handled by applying the mentioned relative measures. On the other hand, young papers that might play an important role cannot be cited very often, as their visibility just begins to raise. Figure 5 illustrates this by plotting citation counts for papers from the CBR community. To overcome this, [Greene *et al.*, 2008] proposes a back-fitting approach. Later, [Greene *et al.*, 2009] applies clustering based on text-similarity in combination with co-citation analysis. We will evaluate these as well as other approaches for handling that crucial issue. Clearly, more research focusing on such timeline effects is needed. Further aspects we will investigate are possible geographical factors (e.g., the differences between American and European conferences) and the filtering of self-citations. It might also be useful to include factors of availability, such as the time at which online publications

are available for certain conferences. A very interesting view is the comparison to methods from computational biology. It can be expected that there exist some parallels, such as communities dying out or surviving, or variants becoming species. One goal is to identify special motifs for certain principles like paradigm shift or (un)healthy communities.

## 4 Conclusion

In this work, we motivated the interesting possibilities and the impact that the analysis of cross-community effects can have. We illustrated phenomena that can be expected and provided a general view on the process to apply. As an outcome, a major roadmap shows the way and challenges that future works will have to take and face. We believe in a very interesting contribution for scientific research in general. Further, we expect the developed methods and gained experiences to be a valuable contribution for analysing community and cross-community effects in other social networks, which are a dominating factor of today's Web.

## References

- [Berners-Lee *et al.*, 2001] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, May 2001.
- [Gaudin and Quigley, 2008] B. Gaudin and A. J. Quigley. Interactive structural clustering of graphs based on multi-representations. In *Int. Conference Information Visualisation (IV'08)*, pages 227–232, 2008.
- [Gmür, 2003] M. Gmür. Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics*, 57(1):27–57, 2003.
- [Granovetter, 1973] M. Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.
- [Greene *et al.*, 2008] D. Greene, J. Freyne, B. Smyth, and P. Cunningham. An Analysis of Research Themes in the CBR Conference Literature. In *European conference on Advances in Case-Based Reasoning (ECCBR'08)*, pages 18–43, 2008.
- [Greene *et al.*, 2009] D. Greene, J. Freyne, B. Smyth, and P. Cunningham. An Analysis of Current Trends in CBR Research Using Multi-View Clustering. Technical report, School of Computer Science and Informatics, University College Dublin, Ireland, 2009.
- [He and Hui, 2002] Y. He and S. C. Hui. Mining a web citation database for author co-citation analysis. *Inf. Process. Manage.*, 38(4):491–508, 2002.
- [Henry and Fekete, 2007] N. Henry and J.-D. Fekete. Nodetrix: a hybrid visualization of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1302–1309, 2007.
- [Kuhn, 1996] Th. S. Kuhn. *The Structure of Scientific Revolutions*. University Of Chicago Press, December 1996.
- [Milgram, 1967] S. Milgram. The small world problem. *Psychology Today*, pages 60–67, 1967.
- [White and Griffith, 1980] H. D. White and B. C. Griffith. Author Co-citation: A Literature Measure of Intellectual Structure. *Journal of the American Society for Information Science*, 32:163–171, 1980.