Citations in Scientific Texts: Do Social Relations Matter?

Béatrice Milard, Ludovic Tanguy

Citations in scientific texts: do social relations matter?

Béatrice Milard *(corresponding author)*
Department of Sociology & LISST (UMR 5193)
University of Toulouse & CNRS
5, allées Antonio Machado
F-31058 Toulouse CEDEX 9
(+33)5 61 50 37 13
milard@univ-tlse2.fr

Ludovic Tanguy
Department of Linguistics & CLLE-ERSS (UMR 5263)
University of Toulouse & CNRS
5, allées Antonio Machado
F-31058 Toulouse CEDEX 9
(+33)5 61 50 36 03
tanguy@univ-tlse2.fr

Abstract: This article presents an investigation of the role of social relations in the writing of scientific articles through the study of in-text citations. Does the fact that the author of an article knows the author whose work he or she cites have an impact on the context of the citation? Since citations are commonly used as criteria for research evaluation, it is important to question their social background to better understand how it impacts textual features. We studied a collection of science articles (N=123) from five disciplines and interviewed their authors (N=84) in order to: 1) identify the social relations between citing and cited authors; and 2) measure the correlation between a set of features related to in-text citations (N=6,956) and the identified social relations. Our pioneering work, mixing sociological and linguistic results, shows that social relations between authors can partly explain the variations of citations in terms of frequency, position and textual context.

Keywords: Citation analysis; social relations; scientific publications; citation content analysis.

1 Introduction

In this paper we propose a mixed approach to the study of citations in scientific articles by considering them from both a social (sociological) and textual (linguistic) point of view. Of course, both disciplines have a long history of studying scientific citations.

Contrasting with practices that consider citations as objective criteria for evaluating research, many sociological works have highlighted their social background, studying the motivations of the citing author (Moravcsik & Murugesan, 1975) or characteristics of the cited one (Bornmann & Daniel, 2008). The scientific community is also a key point for social analysis of citations. Concentrating on wide-coverage data available through bibliometric databases, the science community is seen as a network in which the co-citation links are traces of social groups (Wallace & al. 2012). Some studies based on more local and qualitative aspects have revealed the importance of interpersonal links between authors (Cronin & Shaw, 2002). Even if they insist on intellectual proximities (White, Wellman & Nazer, 2004), these studies are uncorrelated from the texts themselves.

Text analysis of citations consists in identifying the context of in-text citations inside the main text of science articles and extracting features expressing the different choices made by the author (Ding & al., 2014). These features (distribution, position, frequency, surrounding cues, etc.) are then confronted to external characteristics afferent to the cited reference in order to provide the base for an empirical study (Bertin & al., 2016) or to train an automatic classifier (Teufel & al., 2006). Most of the target external characteristics are either subjective (function, opinion or importance, assessed by an annotator or more rarely a posteriori by the author) or loosely related to the citation act (number of citations, age of the reference). Few of these works go further in the analysis of the author's attitude towards his/her citations.

From both sides, most of the studies try to capture the motivation behind a citation, although it is well known that this notion is extremely complex and is influenced by the large number of factors intervening in the writing of an academic article, both internal and external to the article itself. In our opinion, it is more relevant to understand the social and relational context of the citation than the motivation of the authors, which is much more difficult to grasp.

The work we present here proposes to make some new steps in this direction. We deploy citation context analyses aimed at the social relation between citing and cited authors. More precisely, we focus on the level of knowledge between the authors as expressed by the citing author and measure its correlation with a set of features extracted from his/her text. Our hypothesis is that an underlying social relation between authors may impact the inclusion of citations in the text. This social characteristic has not been studied under this scope and we deem it to be an objective contribution to the understanding of citation networks. Our method is therefore mixed and bi-disciplinary, relying on the sociological side on a set of interviews of authors and on the textual side on an automated analysis of the in-text citations extracted from their articles. We base our experiments on 123 articles written by 85 authors from 5 different disciplines, containing a total of 6,656 in-text citations.

By doing so we address the following questions:

- What is the distribution of the social relations between authors across these different disciplines?
- What measurable correlations can be found between these relations and the features of the corresponding in-text citations?
- Do authors treat references differently when citing them according to the social or personal knowledge they have of their authors?

Our results show a complex situation in which it clearly appears that the social relation is significant for a number of the choices made by a citing author, and that citations of references by authors of different levels of knowledge are indeed inserted differently in the text. This confirms that the social context has to be taken into account when analyzing citations, either for research evaluation or for any study based on in-text citations.

This paper is structured as follows. Section 2 reviews the state-of-the-art regarding both social and textual aspects of citations. Our method is presented in Section 3, detailing how we gather our materials on both sides, and the detailed analysis and results are presented and discussed in Section 4.

2 The scientific citation as a social and textual object

Citations have been studied by several disciplines, each addressing specific aspects. Here, we present the current, main issues regarding the citation seen from a sociological and linguistic point of view and how we consider that each one can benefit from the other.

2.1 Social analysis of citations: toward the study of relations between authors

The social dimension of citations has been a longstanding topic of interest for researchers. The motivation for citing another author was the first and principal area of investigation and many studies have attempted to classify the various responses. Scholars who are asked why they cite a reference give several different answers, based on various functions (Chubin & Moitra, 1975; Moravcsik & Murugesan, 1975), including negative ones (MacRoberts & MacRoberts, 1984). Several typologies have been developed that highlight the subtleties and complexity of an author's motivations; they include: to give credit (positive or negative), to alert the reader to a new topic, or simply for information. One in particular, "social consensus" (Brooks, 1985), is presented as an unspecified and vague perception of a consensus within the field. This underlines that motivations to cite are not only a psychological process, but also have a social context. Along these lines, more recent studies have sought to show the influence of the disciplinary context in the choice of references (Harwood, 2009); in particular they emphasize the difference between natural and engineering sciences, and human and social sciences (Larivière, Archambault, Gingras, & Vignola-Gagné, 2006).

Other researchers seek to understand the reasons for citing a particular author. Thus, an increasing number of studies examine the characteristics of the most-cited references, and seek to identify the conditions that favour the citation of a reference (Bornmann & Daniel, 2008; Tahamtan, Safipour Afshar & Ahamdzadeh, 2016). The findings of such studies emphasize the importance of: the number of authors, the length of the article, the journal in which it is published, the journal's ranking, its accessibility, and the scientific field. Other important factors in the choice of citations are the personal characteristics of authors. Scientific renown – often associated with age – is a prime example, first highlighted by Merton who coined the famous term, the "Matthew effect".

Such work claims that an analysis of the individual is sufficient to explain differences in citation rates. Other studies take into account the group, and ask new questions about the social dimension of scientific citations.

Some recent work has questioned the relational dimension or, at least, the place of the individual in his or her peer group. The analysis of citation careers, or the overlap between the references cited by an author and those who cite him or her shows how a researcher's citing circles are gradually

built up (White, 2001; Cronin & Shaw, 2002). Wallace & al. (2012) analysed the link between citations and collaborations and found that citations are a function of the disciplinary and structural conditions in a scientific field.

In these bibliometric studies, the social context is investigated in terms of the traces left in articles themselves (co-authorship, references, journals, etc.). It is well-established, however, that citing is also a process of communication (Cronin, 1984). Reinvestigating the pioneering analysis of invisible colleges (Crane, 1972), citation practices are studied in terms of the social links between researchers – informal communication (Tuire & Erno, 2001) or co-participation in conferences (Zuccala, 2006). These studies consistently show that the congruence between social networks (who knows whom) and citations networks (who cites whom) is incomplete, and underline the need to understand where the differences lie.

In their study based on real-life cases, Johnson and Oppenheim (2007) found that authors' references go beyond their social circles. Baldi (1998) speaks of "intellectual debt" to qualify these cited authors. A longitudinal study of a research group shows that citations reflect not individual strategies in the group but above all intellectual affinities and mutual esteems (White, Wellman, & Nazer, 2004). All these studies show that references in scientific texts are not only the result of an individual choice, but also raise issues about social relations, groups, and institutional and disciplinary partitions.

In this article, we investigate these issues using mixed methods that focus on the relation between the citing and the cited authors. Beyond social motivations or the need to communicate, we argue that focusing on social relations and circles (Simmel, 1955) is a new way to understand the links between citations networks and social networks: do social relations between authors matter for in-text citations' choices?

2.2 Textual analysis of citations: towards automatic classification

In the current section, we present an overview of citation analysis studies that focus on the textual aspects of a citation, and how they can be tested against features such as social relations between authors.

The work presented here belongs to the field of *citation context analysis*, which has seen increased interest in the past few years, mainly thanks to the availability of data (full-text articles in exploitable formats) and efficient Natural Language Processing (NLP) tools; see Bornmann and Daniel (2008), Ding et al. (2014) and Hernández-Alvarez and Gomez (2016) for reviews of the subject.

The interest in citation context analysis is motivated by several objectives, both theoretical and practical:

- *What aspects of the cited work are evoked in the citation?* This can be done through the extraction of keywords from the citation context, which can be used to index the cited reference or to build lexical resources (thesauri) for a disciplinary field (Schneider & Borlund, 2004).
- *What is the function of this citation?* This question is the main target of citation context experiments and studies. It consists of deciding which function (chosen from a pre-established typology) is best-suited to a given citation, based on a number of cues (Teufel & al., 2006). The main applications concern the qualification of citation links in bibliographical databases. There are a number of variations in the area, depending on the number and nature of the functions that are considered. This can range from broad typologies (op. cit.) to narrower distinctions, such as background vs. foreground references (Tanguy & al., 2009).
- *What is the citing author's opinion of the citation?* This question can be seen as a subtopic of the previous one, but with a focus on the identification of positive (praise) and negative (criticism) citations in order to modify bibliometric indicators that are solely based on the number of citations that target a given article/ author/ journal. This sub-question has received much attention, because of its similarity to the task of opinion mining, and thus the availability of resources and tools in applied NLP (Piao & al., 2007).

There are, of course, other considerations that provide incentives and orientations for exploring citation contexts but, ultimately, all automated approaches must exploit the (relevant) characteristics of the citation context. These characteristics need to be easily and automatically extractable (so that the method can be applied to a large number of articles), expressed as simple values (so that they can be processed using standard machine learning classifiers) and – hopefully – relevant to the given task and target characteristics of the citation. Thanks to the large (and growing) number of such studies, these characteristics are now well-known and can be easily listed. We give more detail about the features used in the current study in Section 4 but, in general, they relate to the following aspects: frequency, position, integration and linguistic context.

The exploration and classification of citation contexts is currently popular due to the maturity of NLP tools and the availability of a large amount of raw data, and citation content analyses can be classified as follows:

- *exploratory works* that observe (on a large scale) citation characteristics;
- *contrastive studies* that compare citation behaviours across disciplines, journals, or time periods;
- *classification experiments* that target interpretive characteristics (citation function, opinion, etc.).

The latter category, although the most promising in terms of both knowledge and application, is typically limited by the availability of reliable external characteristics. Citation function is still a very open subject, with a large number of contending typologies—Bornmann & Daniel (2008) list more than a dozen—and the well-known problem of assigning a category even if the annotator is an

expert in the field (Tanguy & al., 2009). Even when acceptable inter-annotator agreement is achieved, there often remains a large number of "unknown", "other" or "neutral" categories (Teufel & al., 2009).

Another target for classification can be the number of citations for a reference as indicated by bibliometric databases (Hu & al. 2013, Ding & al. 2013, Boyack & al. 2018). Fewer studies have focused on labels of their citations by their authors, such as the estimated influence a reference had on their article (Zhu & al. 2015).

In the current study, we take the opportunity to target a more reliable but ignored aspect of citations, while benefiting from the accumulated knowledge and technical know-how of citation content analysis studies.

3 Mixed Methods: interviews, coding and corpus building

In this section we describe the methods used to investigate the possible links between a citing author's social relations with cited authors and the textual characteristic of the citation.

First, semi-structured interviews were conducted with authors from different disciplines regarding one of their published articles. Starting with the list of references, we collected qualitative information about the relations between the interviewee and the cited authors, and used it to define a typology of scientific social circles (Milard, 2014)[1]. To test if these social circles impact the in-text citations, we have selected a number of characteristics that are inherent in the article's text. These characteristics will be presented in detail in Section 4, along with the observed differences.

3.1 Interviews to capture social relations between authors

This first method was an interview, directly linked to a particular article. Each article was selected from the Web of Science published by Thomson Reuters®, one of the most well-known and widely used bibliographic databases; all articles had been published in international, prestigious journals.

Five disciplines or sub-disciplines were investigated: coordination chemistry, molecular biology, mathematics, economics and sociology. They were chosen for their diversity: natural sciences/ social sciences; experimental sciences/ theoretical sciences; and also because of the limited

[1] This article is a part of a broader investigation of scientific articles and their authors. The whole semi-structured interviews focused on the history of each paper and addressed the following points: Where did the idea for the article come from? Who collaborated on the paper and why? Who funded the research? Which authors were cited, and did they know each other? How did the journal evaluate the article? Has the article been cited? (see the interview schedule in Appendix 1).

number of co-authors and references (unlike physics, for example). This choice provided favourable conditions for our method. In the case of articles with multiple authors (80%), the reprint (or corresponding) author was chosen to be interviewed because of his/ her involvement in the publication process. All authors were based in France, which was more convenient for face-to-face interviews. It is possible that this may be a bias in our study as their mother tongue is often (but not always) French. However, all authors have a great practice of English as a professional language, all journals are international and have high standards regarding language. In the end the features we studied are only marginally influenced by the linguistic skill of the authors.

The 123 articles that formed the sample were published between 2004 and 2010 by 84 authors (some authors were interviewed about two of their articles). The selection criteria for the corpus were as follows: the authors had to 1) be tenured; 2) have published at least two articles as corresponding author within a period of 5 years; 3) be a member of a recognized laboratory in their field; 4) have accepted the interview. To have fruitful interviews, we selected articles with at least 15 references. In the end, we have between 25 and 30 chemistry, economics, biology and mathematics articles and 11 sociology articles. We were targeting an equal number of articles in the five disciplines, but the constraints were too strict for sociology (especially the fact that a large number of publications are in French).

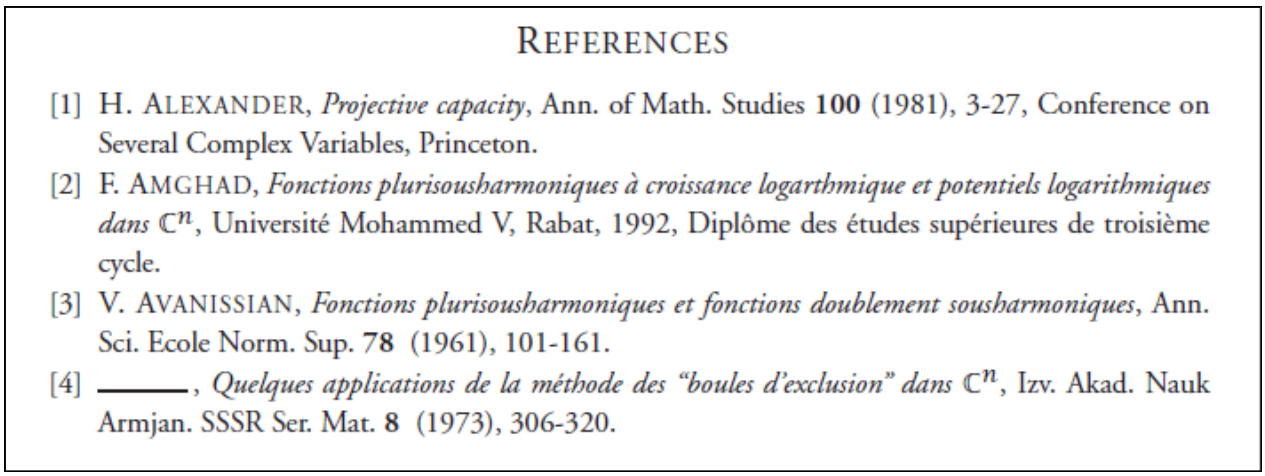
Interviews were conducted in 2012 and 2013, with the exception of those in the discipline of chemistry, which were conducted before 2010. They were conducted by three different interviewers, mainly in the authors' offices, and lasted between one and three hours each.

During the interview, the researcher under study answered the following question regarding each author listed in the bibliography[2]: "Can you characterise the type of relationship you have with this person whose work is cited in the list of references, and if you do not know him/her personally, what do you know about him/her?". Based on the transcribed interviews, more than 8000 qualitative answers were then hand-coded by several coders, who classified the relationships. All data was gathered in an ad-hoc relational database. As in previous work (Milard, 2014), this classification was based on the Simmelian theory of social circles (Simmel, 1955), applied to the scientific world. A social circle is defined as a group of individuals who have common characteristics or activities, based on mutual recognition. The link can be positive or negative, but the criterion that is taken into account is objective, and based on their relational experience: Do they know each other? Have they met? Do they work together?

An example will make this clearer. Figure 1 shows the first items in the target paper's list of references, while Figure 2 contains an extract from the corresponding transcribed interview.

---

[2] For time efficiency, and because the main target of this part of the interview was cited authors, we followed the list of references and not the flow of citations in the body of the article's text.



References

[1] H. Alexander, *Projective capacity*, Ann. of Math. Studies **100** (1981), 3-27, Conference on Several Complex Variables, Princeton.

[2] F. Amghad, *Fonctions plurisousharmoniques à croissance logarthmique et potentiels logarithmiques dans* $\mathbb{C}^n$, Université Mohammed V, Rabat, 1992, Diplôme des études supérieures de troisième cycle.

[3] V. Avanissian, *Fonctions plurisousharmoniques et fonctions doublement sousharmoniques*, Ann. Sci. Ecole Norm. Sup. **78** (1961), 101-161.

[4] ———, *Quelques applications de la méthode des "boules d'exclusion" dans* $\mathbb{C}^n$, Izv. Akad. Nauk Armjan. SSSR Ser. Mat. **8** (1973), 306-320.

Figure 1: Bibliographical references in a mathematics article (extract)

Interviewed author (IA): *So, Alexander* **[ref 1]**, *I know him by name, I have never met him. Amghad* **[ref 2],** *in fact she is one of my former students, with whom I worked in Rabat. Avanissian* **[ref 3 and 4],** *he is a French mathematician from Strasbourg that I met once, yes...*

Sociologist (S): *You talked to him or ...?*

IA: *Yes, yes, I talked to him, we exchanged emails… He works on this subject because he was interested too…*

S: *And you've met him during a congress or...*

IA: *Yes, during a congress. Then, Bedford and Taylor* **[ref 5 and 6],** *in fact, it is really a fundamental reference, because they were the pioneers of this theory in the 80's. And I was lucky at the time to have met them at the beginning. This is what oriented me to this topic. Bedford and Taylor is really a key reference. […]*

IA : *Demailly, Demailly* **[ref 14 and 15],** *he is also a great French mathematician, whom I know very well. We wrote an article together too.*

S : *He's in Toulouse?*

IA : *No, he's in Grenoble.*

[…]

IA : *Siciak* **[ref 32 and 33],** *he is a Polish mathematician, whom I know very well. He has come several times here. We know each other very well, we work well together.*

[…]

IA: *and the others* **[ref 37 to 40]** *are my own articles.*

Figure 2: Interview of the main author of the mathematics article (extract)

The interviews identified six types of relationships between the citing (interviewed) author and the cited author. They are detailed below in order of decreasing closeness to the interviewed author.

The first type of relationship concerns self-citations (*1_Self* hereafter), i.e., when the author cites his or her own work, as can be seen in the example of references 37 to 40 in Figure 2.

The second and third categories concerned either members of the same team or laboratory (*2_sameLab*), or collaborators and friends (*3_Collab/Fri*end) of the interviewed author. Both colleagues and students were considered as a member of the team (*2_SameLab*), as seen in reference 2 in Figure 2; the decisive criterion is that the parties met in an institutional setting.

Collaboration is the most frequent example of a non-institutional relationship (*3_Collab/Friend*), but close, informal (friendly) links also exist. This is the case of the author cited in references 32 and 33 in Figure 2, who the interviewee knows well, although they have never published a paper together.

The fourth category concerns cases where the interviewee has only met the author briefly, or where they have both attended a conference or meeting (*4_AlreadyMet*). This is the case for references 4 and 5 in Figure 2.

Category five concerns authors who are known only by name (*5_KnownByName*). These authors are often famous or frequently referenced in the literature. This applies when the interviewee can give at least some information about the author (position, notable achievements, etc.).

Finally, there are a number of authors who are unknown to the interviewee, i.e. they can provide no information at all (*6_Unknown*). They may be unknown because of their professional status (doctoral student, technician, etc.), because they are spatially distant from the citing author (for example Chinese colleagues), because they are from another time period (for example an author from the 1930s), or from another discipline (for example a chemist citing a physicist).

For more details on the types of relationship, see Milard (2014).

3.2 Corpus presentation

In this section we present the corpus of articles, and describe how it was processed. Our corpus consisted of 123 articles whose origin is described in Section 3.1.

The full text of each article was extracted from files in PDF format (based on the published versions provided by the interviewed authors) and cleaned up. The corpus contained an overall total of 1.1 million tokens. The breakdown, per discipline, is indicated in Table 1.

Discipline	Nb. of articles	Nb. of words (per article)		Nb. of references (per article)		Nb. of citations (per article)	
Biology	25	193,146	(7,726)	981	(39)	1672	(67)
Chemistry	30	187,369	(6,246)	1373	(46)	1995	(66)
Mathematics	26	376,868	(14,496)	628	(24)	1205	(46)
Economics	31	307,346	(9,914)	1002	(32)	1486	(48)
Sociology	11	112,798	(10,254)	418	(38)	598	(54)
ALL	**123**	**1,177,547**	**(9,573)**	**4,401**	**(36)**	**6,956**	**(57)**

TABLE 1: CORPUS SIZE AS A FUNCTION OF DISCIPLINE

While there are large differences in terms of article length between the five disciplines (with mathematics and social sciences articles being much longer than biology and chemistry), the average number of citations per paper shows a notable homogeneity.

In-text citations were explicitly marked up (using an ad-hoc XML schema), and allocated a number that was cross-referenced to the information obtained from the interviews with authors, and details of the reference. We were therefore able to standardise the different citation styles used in different disciplines and/ or journals. The same tag was used to replace full-name, Harvard-style citations (e.g. "*(Smith 2009)*"), abridged names (e.g. "*[SMI09]*"), Vancouver-style numbers (e.g. "*[1]*", "*(1)*" or footnote/ endnotes. In total, 6956 citations of 4401 references were identified.

This preparation was necessary for the subsequent automatic processing step, which is presented in the next section.

4 Analysis and results

In this section, we present the main results of our study. We analyse the citation characteristics of our corpus, and compare them to the social relations between the citing and cited authors. In cases where there are multiple authors, there may be several types of relationship within the same reference. In such cases, we chose to retain the closest relationship, as it tends to prevail over the others. The aim was to identify any correlations, and the underlying question could be rephrased as "does the level of social relation influence the action of citation?". A further consideration is inter-discipline variation.

As a preliminary result, Figure 3 presents the relations between citing and cited authors, for all references, according to the disciplines.

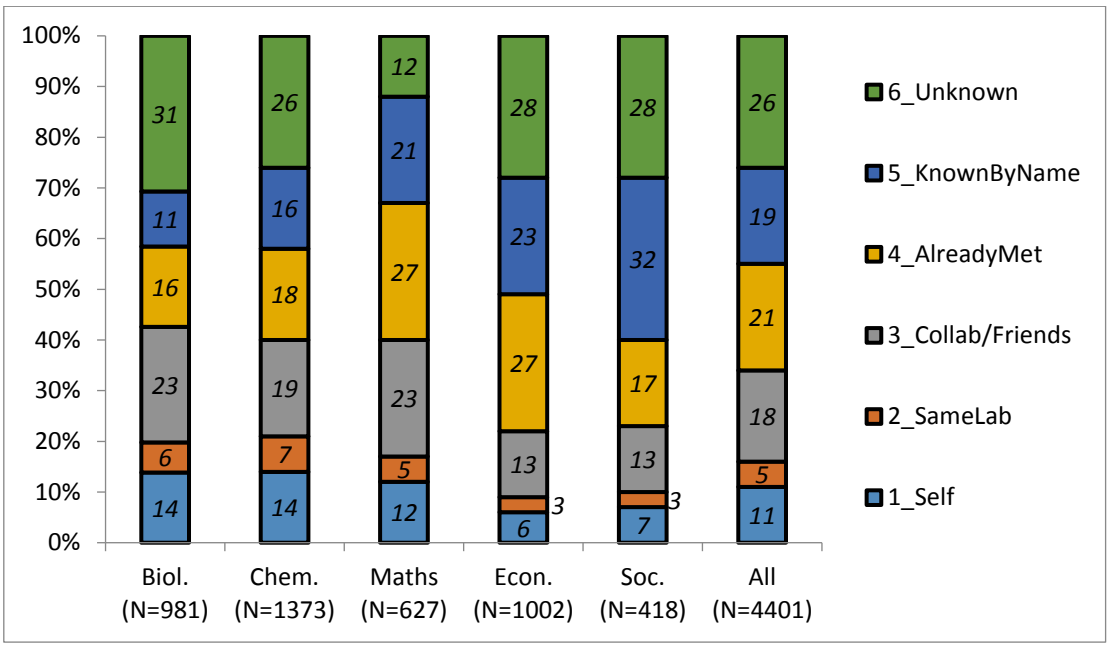


FIGURE 3: REFERENCES AS A FUNCTION OF SOCIAL RELATION AND DISCIPLINE (CLOSEST RELATION)

To the best of our knowledge, there are no similar, empirical studies of these sorts of relations. The only comparable results appear to concern self-citations. These studies are popular, as self-citations can easily be identified in a reference list without having to interview the authors. The most complete study on the subject (Glänzel & Thijs, 2004) presents similar results in terms of distribution: self-citation is more prevalent in biology, chemistry and mathematics (around 13%) than in sociology and economics (around 7%). Overall, the profile for biology resembles that of chemistry. Likewise, economics and sociology are somewhat similar, although the percentage of known authors is higher in economics. Mathematicians have the distinction of citing the fewest unknown authors in their bibliography.

However, these initial results do not take into account the details concerning the insertion of a reference in a text (i.e. the citation). To understand the links between social relations and the content of the text, it is essential to look at the different ways citations are used.

We studied a number of characteristics that are known to vary across citations, and can be seen as an indication of the attitude of the citing author to the cited author. We present these features in order of increasing abstraction and subjectivity and, incidentally, processing complexity. As we will see, most features can be related to a (as yet imprecise) notion of ‘centrality’ or ‘importance’ of a citation. Our hypotheses are: that it is not neutral for an author to cite the same reference several times (Section 4.1); that a citation in the introduction is different to a citation in the methodology (Section 4.2); and that an isolated citation is different to a citation in a cluster of several other references (Section 4.3); this choice is based on the work of, for example, Tanguy & al. (2009). The text that surrounds a citation is also an indicator of the author’s attitude towards it although, it is, of course, the most complicated information to use (Section 4.4).

4.1 Citation distribution and frequency

The first set of characteristics we consider is simply the distribution of social relations across all the citations in our corpus, which naturally leads to a cross-discipline comparison and an analysis of the number of citations for a given reference.

4.1.1 Number of citations per discipline and type of social relation

The first point to note is that this analysis focuses on *in-text citations*, and not the list of references. Therefore, the distribution is slightly different to that observed in Section 3 (Figure 3). The number of citations for each discipline and each type of social relation is indicated in Figure 4.

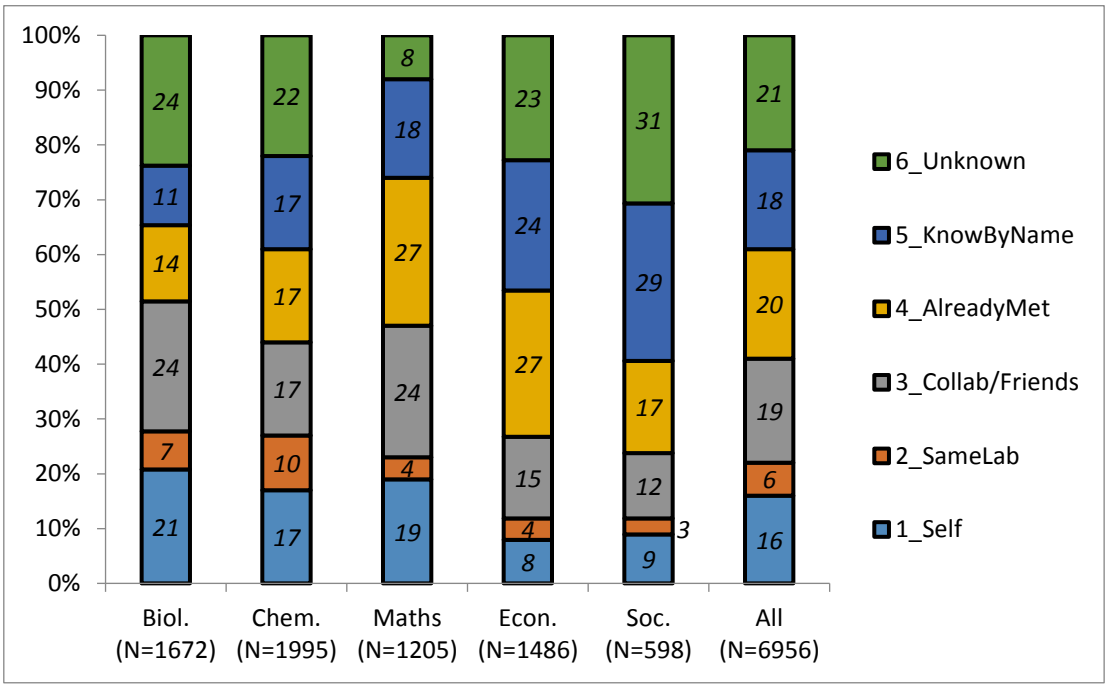


FIGURE 4: IN-TEXT CITATIONS AS A FUNCTION OF DISCIPLINE AND SOCIAL RELATION

The overall trends are similar to those observed for the list of references. The only notable differences concern an increased ratio of self-citations (16% compared to 11%) and a lower ratio of unknown authors (21% compared to 26%), which indicates that the number of citations per reference varies, as we will see in the next section.

When we compare disciplines, we see that the overall level of relationship with cited authors is higher for mathematics, with a higher proportion of *1_Self* and *3_Collab/Friend*, and very few citations of unknown authors. This trend is inverted for the social sciences, with sociology at the other end of the spectrum. Experimental sciences are similar to mathematics, but with a much higher rate of unknown authors, and even higher rates of self-citation. This confirms the observations of Snyder & Bonzi (1998) on self-citations, which were found to be significantly more frequent in experimental sciences.

4.1.2 Number of citations per reference

As noted above, the difference in the distribution of references compared to citations indicates that there are differences in the number of times a given reference is cited. This simple to measure feature has been frequently studied in citation analysis. (Zhu & al. 2015) have measured that the number of times a reference is cited in an article is correlated to the citing author's estimation of the reference's influence on their work, while (Boyack & al. 2018) found that references cited only once in an article are generally highly cited according to bibliometric databases.

The average number of citations for a given reference in our corpus is 1.58 (±0.04, 95 % CI). Table 2 shows the detailed results as a function of discipline and social relation.

Discipline	1_Self	2_Same Lab	3_Collab/ Friend	4_Already Met	5_Known ByName	6_Unknown	Total
Biology	2.66	2.06	1.75	1.47	1.76	1.29	**1.70**
Chemistry	1.75	2.00	1.34	1.38	1.49	1.25	**1.45**
Maths	2.94	1.59	2.03	1.88	1.66	1.36	**1.92**
Economics	1.86	1.68	1.69	1.47	1.58	1.22	**1.48**
Sociology	1.80	1.15	1.38	1.38	1.30	1.56	**1.43**
All	**2.20**	**1.86**	**1.63**	**1.52**	**1.54**	**1.29**	**1.58**

TABLE 2: AVERAGE NUMBER OF CITATIONS PER REFERENCE

The colours of the cells in Table 2 show, for each discipline, the higher (in green) and lower (in red) values, i.e. the relative variation in the numbers of citations per social relation for each discipline, and for all articles. The overall picture thus seems to indicate that, with a few exceptions, references that correspond to stronger social relations are cited more frequently.

However, over all five disciplines, there are in fact significant[3] differences across social relation types ($F=33.2$, $df=5,4395$, $p<0.001$). A closer look indicates that *1_Self* citations are cited most frequently: more than two citations per reference. This is significantly ($p<0.05$ for all categories) higher than all other types. Although no differences are found for *2_SameLab* and *3_Friend/Colleague*, the citation rate is significantly higher than for the other three categories ($p<0.05$). And, finally, *6_Unknown* citations are significantly lower than all others ($p<0.05$). No significant differences are found for Levels 4 (*AlreadyMet*) and 5 (*KnownByName*).

Several hypotheses can be formulated to explain the behaviour observed for self-citations, such as the fact that they are frequently used to indicate the follow-up of earlier work by the authors. The low repetition rate of *6_Unknown* authors could indicate that the function of these references in an article is less central, and that they serve to give background or methodological information.

[3] ANOVA and post-hoc Tukey HSD tests were used, with a significance threshold of 0.05.

The results highlight differences between disciplines, notably mathematics and biology have a higher overall repetition rate than the other three disciplines. It is interesting to note that this finding is independent of the article's length, or even the total number of citations per article which show significant differences between these disciplines (see Table 1).

The analysis of citation frequency against relations for each individual discipline found significant differences in all cases, with the exception of sociology (biology: $F=18.3$, $df=5,975$, $p<0.001$; chemistry: $F=15.0$, $df=5,1367$, $p<0.001$; maths: $F=5.5$, $df=5,621$, $p<0.001$; economics: $F=5.1$, $df=5,996$, $p<0.001$; sociology: $F=1.3$, $df=5,412$, $p>0.05$). However, only the higher frequency of self-citations and lower frequency of unknown citations can be confirmed for each of the five subsets ($p<0.05$ for all comparisons).

4.2 Position of citations

Another important characteristic of a citation that can be easily measured is the position in which it appears in the text. Several large-scale analyses have studied the distribution of citations in an article, based on their relative position. (Hu & al. 2013) found that references cited in the first parts of an article have globally higher citation rates. (Bertin & al. 2016) found invariant schemes of distribution of citation in the position of citations, but (Boyack & al. 2018) found important differences across disciplines. The underlying hypothesis in this study and in the current paper is that the position of a given citation is influenced by its relative importance or function.

4.2.1 Absolute position (offset)

For each citation, we measured its absolute position in the text, measured as the character offset of the corresponding XML tag. These values were normalised in such a way that absolute position values ranged from 0 (the citation appears at the very beginning of the body of the text) to 100 (it appears at the very end).

Figure 5 shows the relative position of citations as a function of discipline and social relation.

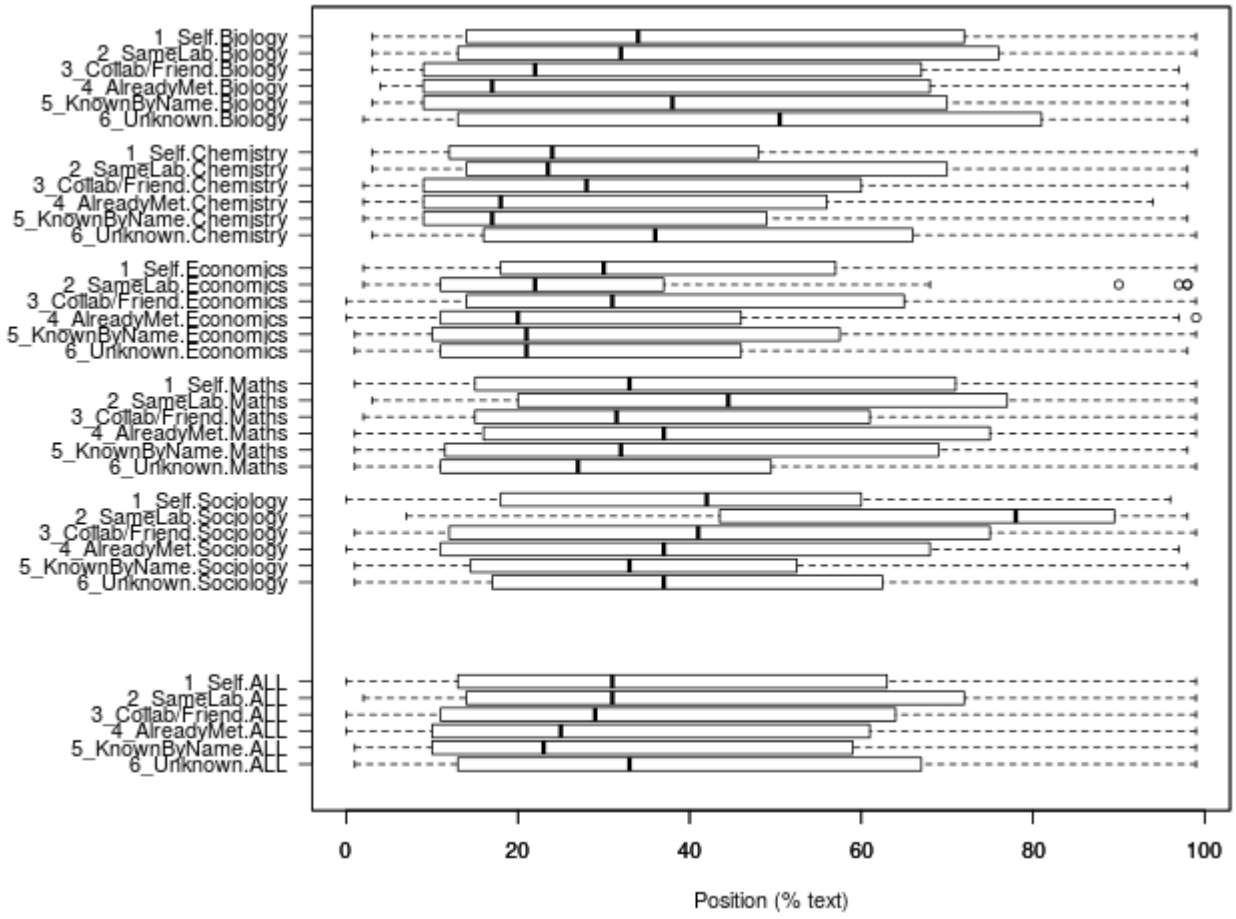


FIGURE 5: OFFSET OF IN-TEXT CITATIONS AS A FUNCTION OF DISCIPLINE AND SOCIAL RELATION

Figure 5 shows that most citations occur in the first parts of an article, but rarely at the very beginning. The median position of a citation in the overall corpus is 29%, which corresponds exactly with the findings of Zhu & al. (2013), who analysed a homogenous corpus of articles from the *Journal of Informetrics*.

There are significant variations in global positions across disciplines ($F=23.7$, $df=4,6951$, $p<0.001$). More precisely, in chemistry and economics citations appear significantly earlier in the text than for the other three disciplines ($p<0.05$). This is somewhat surprising as the articles from these two disciplines have different lengths and structures.

Variation in position as a function of social relation is also significant ($F=7.6$, $df=5,6950$, $p<0.001$) but more elusive. The only significant difference, found using Tukey's HSD test, concerns citations in the *4_AlreadyMet* and *5_KnownByName* categories, which both appear earlier than *6_Unknown* citations ($p<0.05$). However, the lower set of boxplots shown in Figure 5 indicate that, with the exception of the *6_Unknown* category, the better-known the author, the later the citation appears in the text.

When analysing each discipline individually, we found that although significant differences could be identified, it was difficult to define a common trend that was stable across disciplines. This difficulty

is partly caused by wide variation in article structure, which is not satisfactorily represented by the simple relative position. Therefore, in the next section we propose an alternative approach.

4.2.2 Position of citation (sections)

The IMRaD (Introduction, Methods, Results and Discussion) prototypical structure has been thoroughly described in the literature, as it is prescribed (and followed) for most science articles (Swales, 1990). Several recent studies of citation behaviour have focused on the section in which a citation appears, whether to study linguistic variations in the context (Bertin & Atanassova, 2014), the age of a citation (Bertin & al., 2016), or its function (Teufel, 2010). In order to explore this aspect, we manually tagged each section heading in our corpus according to the IMRaD categories. The overall results are presented in Table 3. The IMRaD column indicates the number of articles that follow the IMRaD structure (not necessarily in this order, as the method/ experimental sections appear at the very end in about 50 % of biology and chemistry articles). Articles that do not follow this structure either have no section headings whatsoever, or use thematic names (e.g. "*Geometric Currents",* "*The Abstract Existence Result",* "*Projective Masses"*). The last two columns indicate the number of articles that have an explicit Introduction and/or Conclusion.

Discipline	Nb of Articles	Full IMRaD	With Introduction	With Conclusion
Biology	25	25 (100%)	25 (100%)	6 (24%)
Chemistry	30	22 (73%)	22 (73%)	17 (57%)
Maths	26	0 (0%)	24 (92%)	2 (8%)
Economics	31	1 (3%)	31 (100%)	31 (100%)
Sociology	11	1 (9%)	9 (82%)	8 (73%)
All	**123**	**49 (40%)**	**111 (90%)**	**64 (52%)**

TABLE 3: DISTRIBUTION OF IMRAD SECTIONS IN THE CORPUS

The results presented in Table 3 show that the IMRaD structure does not apply in mathematics or social sciences, and is limited to experimental sciences (as expected). A more surprising finding is the lack of a conclusion, including in the absence of the IMRaD structure. The exception is social sciences articles, which systematically include it. This information provides the input for an analysis of the distribution of the social relationship as a function of the section in which the citation appears.

We first focused on the two disciplines for which the IMRaD structure is relevant (biology and chemistry), and test whether there are significant differences in the way citations are distributed across sections as a function of the relation with the cited author. We considered only the 47 (25 biology and 22 chemistry) articles with a full IMRaD structure. We merged the "Results" and "Discussion" sections, as most authors group them into a "Results and Discussion" section. Table 4 shows adjusted Pearson residuals for the Introduction (I), Methods (M), Results and Discussion (RD) and Conclusion (C) sections. Values are highlighted according to their sign (green for positive,

red for negative) and absolute value. Brighter colours are used for residuals with an absolute value greater than two, as it is commonly considered as a threshold for significance (Agresti, 2013).

Section	1_Self	2_SameLab	3_Collab/ Friend	4_AlreadyMet	5_Known ByName	6_Unknown
I	-3.37	0.52	3.54	4.89	4.68	-7.41
M	3.99	1.10	-6.45	-4.29	-6.09	9.32
RD	0.30	-1.43	1.40	-1.59	0.10	0.21
C	0.74	0.39	-1.14	-0.75	-1.39	1.81

TABLE 4: IMRAD SECTIONS AND SOCIAL RELATIONS IN BIOLOGY AND CHEMISTRY ARTICLES (ADJUSTED PEARSON RESIDUALS)

The most salient results concern the citations in the *6_Unknown* category, which appear most often in the Method section and, to a lesser extent, in the Conclusion, but are rarely found in the Introduction. This is understandable, as many citations used in these sections refer to standard methods, for which knowledge of the authors, or a close relationship with them, is mostly incidental.

Significant differences are also found in the distribution of *1_Self* citations, being rare in the Introduction, but more frequent in the Method section.

As for the other categories, we can see that most of the other categories follow the same distribution: *3_Collab/Friend*, *4_AlreadyMet* and *5_KnownByName* citations are more frequent in the Introduction and rare in the Method. No significant variation is found for the *2_SameLab* category.

Examining variation across all five disciplines requires us to abandon the IMRaD structure, and to target the lowest common denominator regarding the sections found in articles. We thus measured the association between a citation's presence in the Introduction and its category. More precisely, we compared the distribution (across categories) of all citations in the Introductions we found to the global distribution of categories. We did this for both the whole corpus, and separately for each discipline. Adjusted Pearson residuals are shown in Table 5; here again, absolute values above two have been highlighted.

Discipline	1_Self	2_SameLab	3_Collab/ Friend	4_AlreadyMet	5_Known ByName	6_Unknown
Biology	-3.19	0.98	3.57	3.81	0.58	-3.23
Chemlstry	-1.96	0.19	0.44	2.31	6.04	-5.86
Maths	0.68	1.04	-0.84	-0.45	-0.45	0.95
Economics	-1.79	-0.79	-2.44	2.94	0.80	-0.40
Sociology	1.72	-0.86	2.31	0.46	-1.48	-1.34
All	**-1.62**	**0.21**	**0.77**	**3.46**	**1.83**	**-4.05**

TABLE 5: DISTRIBUTION OF SOCIAL RELATIONS FOR CITATIONS IN INTRODUCTIONS (ADJUSTED PEARSON RESIDUALS) AS A FUNCTION OF DISCIPLINE

For biology and chemistry, the effects noted in the previous table appear once again. For the other disciplines, however, the results differ. No significant variation can be observed for mathematics; in economics, the Introduction features citations of less-well-known authors; sociology articles contain more self-citations, and more citations of the work of collaborators and friends.

We then applied the same procedure to the Conclusions: the results are presented in Table 6.

Discipline	1_Self	2_SameLab	3_Collab/ Friend	4_AlreadyMet	5_Known ByName	6_Unknown
Biology	-1.33	-0.82	-2.72	-1.14	-0.71	5.47
Chemlstry	2.23	1.40	0.72	-0.27	-1.22	-2.07
Maths	0.31	-0.31	-1.13	0.36	0.97	-0.73
Economics	-0.57	-0.68	0.12	-0.76	3.10	-1.74
Sociology	-1.74	3.98	-0.31	2.59	0.28	-2.43
All	**-1.38**	**1.70**	**-1.63**	**0.04**	**2.26**	**0.12**

TABLE 6: DISTRIBUTION OF SOCIAL RELATIONS FOR CITATIONS IN CONCLUSIONS (ADJUSTED PEARSON RESIDUALS) AS A FUNCTION OF DISCIPLINE

This detailed measure shows that there are differences between biology and chemistry, which could not be seen in the first study (Table 6). Specifically, the conclusions of chemistry articles favour self-citations, although this is the only discipline for which such behaviour is observable. Biology has a high frequency of unknown citations in the Conclusion, unlike all of the other disciplines.

These differences are less significant than for introductions. This is mostly due to the lower frequency of articles with a conclusion, notably in mathematics where only 8% of articles have an explicit Conclusion.

To sum up, while clear trends can be identified regarding self-citation and the citation of unknown authors, few significant variations can be observed for the other categories, and no general trend can be identified across disciplines (note that the contents of Tables 5 and 6 were computed independently based on the corresponding subsets of our data).

4.3 Citation clusters

In order to study the citation context in more detail (i.e. the textual content of the sentence in which a citation occurs), we first need to explore the heterogeneity of these contexts with respect to the social relation. We consider that citations often appear in clusters and that, in these cases, the citing authors treat them as equivalent.

In the following, we consider that two citations appear in the same cluster if they meet strict criteria. This can be one of the following cases:

1. both citations appear in the same footnote or endnote, with no intervening text;
2. both citations are grouped within the same parenthesis, such as "*(7, 8, 9)*";

3. both citations appear in the same sentence with no intervening text, such as "*see MacDonald, 1997; Michael et al., 1997; Pippenger and Goering, 1998; Baum et al., 2001; Taylor et al., 2001*".

If there is a textual element between two citations, they are considered to be part of separate clusters, illustrated by "*[…] proposed by Cass (1972) and, concerning OLG economies, by Shell (1971) [...]*".

We extracted all such clusters from our corpus, retaining the social relation type for each citation. Our objective in this part of the study was to measure the distribution and combination of social relations in these clusters, where the citing authors considered the citations as equivalent.

Table 7 shows several statistics related to the distribution of citations in clusters.

Discipline	Citations in clusters	Avg. number of clusters	Max. cluster size	Avg. cluster size
Biology	47%	12.5	7	2.5
Chemistry	61%	12.0	16	3.5
Maths	27%	4.7	11	2.5
Economics	38%	7.2	7	2.7
Sociology	18%	4.8	6	2.6
All	41%	8.7	16	2.8

TABLE 7: DISTRIBUTION OF CLUSTERED CITATIONS AS A FUNCTION OF DISCIPLINE

There is an important variation across disciplines in the clustering of citations. Chemistry articles contain numerous and large clusters, covering a majority of the citations; most of them are in fact footnotes containing nothing but a list of references. Biology follows a similar scheme, but with a larger number of smaller clusters (citations are mostly inline). On the other side, Sociology and Maths authors scarcely regroup their citations in clusters, with a large majority of citations appearing isolated. Economics articles show a profile very close to the average on these aspects.

We then measured the distribution of social relations inside these clusters and for isolated citations. The order in which citations appear in a given cluster was not taken into account, as it may be conditioned by arbitrary criteria (such as alphabetical or chronological order).

First, we computed the number of citations of each type that appear isolated, then the number of clusters in which a given pair of social relation types appeared. The overall results are presented in Table 8. For example, we found 1006 isolated self-citations, 239 clusters where two or more self-citations were listed, 16 clusters where a self-citation co-occurred with a *2_SameLab* citation, and so on.

Appears with	Isolated	1_Self	2_SameLab	3_Collab/ Friend	4_Already Met	5_Known ByName	6_Unknown
1_Self	1006	239	16	89	89	59	54
2_SameLab	73		7	14	16	17	6
3_Collab/Friend	517			118	119	58	76
4_AlreadyMet	781				155	137	128
5_KnownByName	707					130	122
6_Unknown	825						178

TABLE 8: COMBINATIONS OF SOCIAL RELATION TYPES IN CITATION CLUSTERS (EXPRESSED AS THE NUMBER OF CLUSTERS OR ISOLATED CITATIONS)

We then performed a Monte-Carlo simulation in order to estimate the distribution of social relation types in clusters when all citations are randomly distributed. To do this, for each article in our collection, we ran 1000 random shuffles of all citations, preserving the original number and size of each citation cluster (including isolated citations). This gave us an expected distribution, which we used as a baseline for the identification of significant variations.

Table 9 shows adjusted Pearson residuals measured when comparing the observed distribution (in Table 8) with expected values. Absolute residuals above two are highlighted (green indicates an observed frequency higher than expected, and red lower). It appears that self-citations are most likely to be isolated or associated with other self-citations, but appear very rarely in association with other social relation types. The same principle applies to *6_Unknown* citations, which are not found in combination with other types (especially self-citations). Mid-range types (*3_Collab/Friend* and *4_AlreadyMet*) tend to appear in a group (i.e. not isolated) and there is a good level of compatibility between them. Globally, although homogeneity is observed for all types of clustered citations, the phenomenon seems to be stronger for citations with a low level of knowledge of the cited authors. Another observation is the limited compatibility of citations of similar types, as shown in the second diagonal of Figure 9 (1 & 2, 2 & 3 etc.).

Appears w/	Isolated	1_Self	2_SameLab	3_Collab/ Friend	4_Already Met	5_Known ByName	6_Unknown
1_Self	3.37	11.89	-0.68	-3.53	-8.47	-8.80	-15.42
2_SameLab	0.06		12.27	-0.67	-3.31	-10.89	-10.03
3_Collab/Friend	-2.89			24.02	1.70	-0.19	-3.27
4_AlreadyMet	-2.29				9.26	-0.48	-4.32
5_KnownByName	-0.49					21.77	1.70
6_Unknown	2.38						13.38

TABLE 9: ADJUSTED PEARSON RESIDUALS OF SOCIAL RELATION TYPES MIXED IN CITATION CLUSTERS (ALL DISCIPLINES)

Figure 6 is a rough outline of the residuals for each discipline, following the same structure as Table 9 (the first column is for isolated citations, other columns and rows correspond to relation types indicated by their numbers). Exact values are not reported, but lighter colours indicate either positive or negative values under a threshold of two. Similar tendencies can be observed for each subset of the data, although the effects are less visible for disciplines that favour isolated citations, such as maths and sociology.

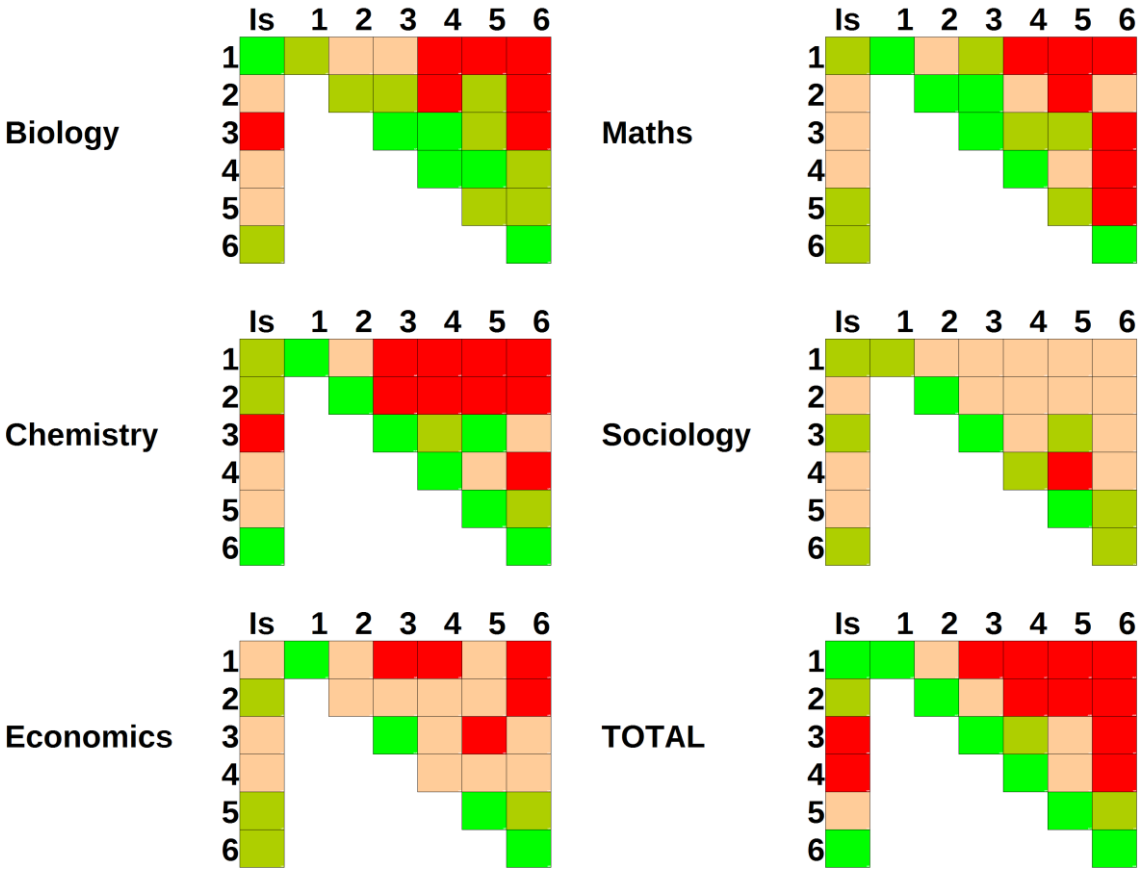


FIGURE 6: ADJUSTED PEARSON RESIDUALS FOR COMBINED SOCIAL RELATION TYPES IN CITATION CLUSTERS FOR THE FIVE DISCIPLINES AND THE OVERALL CORPUS.

4.4 Textual content of the citation context

The last aspect we considered in the evaluation of variation in social relations was the context, i.e. the text surrounding the citation. This aspect of a citation is typically useful in order to identify its function (Teufel, 2006).

The work we describe here represents an initial attempt to identify whether there are any differences in the linguistic aspects of a citation, indicative of differences in the attitude of the citing author to the different types of citation.

To define the context of a citation, we adopt what Ritchie et al. (2008) call the *3sentupto* method, which consists in extracting the sentence the citation appears in, together with the preceding and following sentences, but truncating the context when another citation is found to the left or right of the target citation.

We restricted our study to isolated citations and homogeneous citation clusters (i.e. clusters in which all citations belong to the same social relation category). As noted above, clustered citations cannot be distinguished from one another, and within-group heterogeneity leads to contradictory information. In addition, large clusters would dominate if each citation in the cluster was taken into

account. Whatever the homogeneous cluster's size, we thus considered it as only one instance of the corresponding social relation. This led to the identification of 4,356 citation "spots", characterised by a single social relation, for which we extracted the contexts.

We then trained a supervised logistic regression classifier (with a cut-off of 5) to categorize the context according to the social relation. We used standard features to represent the context (word lemmas unigrams, bigrams and trigrams; bigrams of lemmas+POS tag; trigrams of characters). Each feature was limited to a binary value (present/ absent from the context). The classifier was evaluated based on 10-fold random cross-validation.

The resulting accuracy was 43% and the micro-average F1-measure was 0.43. Detailed scores as a function of social relations are indicated in Table 10.

Social relation	Precision	Recall	F1
1_Self	0.49	0.48	0.48
2_SameLab	0.23	0.39	0.28
3_Collab/Friend	0.43	0.43	0.43
4_AlreadyMet	0.42	0.40	0.41
5_KnownByName	0.39	0.40	0.40
6_Unknown	0.49	0.46	0.47

TABLE 10: SOCIAL RELATION SCORES FOR THE TRAINED CLASSIFIER (10-FOLD CROSS-VALIDATION)

It is clear that different social relations influence the efficiency of the classifier. Self-citations and unknown citations were identified most easily. The remaining levels have less accurate, but similar scores. Similar values were found when we processed each discipline subset separately.

Next, we investigated the best predictor features for each class, based on their weights in the trained model. Specifically, we selected word and word n-grams (as character n-grams cannot be interpreted easily), and manually examined some of the contexts to obtain an insight into linguistic phenomena.

For the *1_Self* category, the best linguistic features related to temporal expressions such as "*previously*" and "*recently*" (as in "*as previously shown*"), and first-person pronouns and possessives ("*we*" / "*us*" / "*our*"). This confirms that most self-citations explicitly refer to previous work by the authors.

For the *6_Unknown* category some features related to methodological aspects ("*use*", "*parameter*", "*software*", "*database*" etc.) and expressions such as "*e.g.*" and "*for example*", indicating that the cited reference is just one out of several other possible choices.

For the other categories, although some features were identified they could not be interpreted as associated with the target class. This can be explained by the low scores achieved by the classifier.

Finally, the confusion matrix of the classifier is shown in Table 11.

Actual \ Predicted	1_Self	2_SameLab	3_Collab/ Friend	4_Already Met	5_Known ByName	6_Unknown
1_Self	48%	17%	15%	11%	8%	10%
2_SameLab	6%	39%	3%	5%	3%	5%
3_Collab/Friend	14%	12%	43%	14%	12%	10%
4_AlreadyMet	11%	14%	16%	40%	19%	13%
5_KnownByName	9%	7%	11%	16%	40%	16%
6_Unknown	11%	12%	12%	14%	18%	46%

TABLE 11: CONFUSION MATRIX OF THE TRAINED CLASSIFIER (10-FOLD CROSS-VALIDATION), NORMALISED PER COLUMN

It appears that, like citation cluster homogeneity, confusion between categories is higher when relation types are similar, and this is especially true for the lower end of the social relation spectrum. In other words, citations of similar levels (4 and 5, 5 and 6) are more difficult to discriminate. This indicates that the citation context is impacted by the social relation, although its discriminative characteristics remain to be identified more precisely.

5 Conclusion

In this article, we developed a new approach to exploring scientific citations. We focused on the relationships between citing and cited authors and we developed several analyses concerning the impact of these social links on scientific texts.

This led to the clear identification of several significant characteristics of self-citations and citations of unknown authors.

Self-citations are more common in the experimental sciences and mathematics than in the social sciences. They are repeated more often in all sections of the article; however, for articles that have an IMRaD structure they appear less frequently than other types in introductions and conclusions, but more often in methodology sections. They are usually isolated, or appear alongside other self-citations, and have distinctive contextual elements (personal pronouns, past adverbials).

Unknown citations are the most common type in all disciplines, except mathematics. They are rarely repeated and rarely appear in introductions. They are generally isolated or appear with other unknown citations and, like self-citations, have distinctive contextual elements (vocabulary associated with the methodology).

As for the intermediate categories (i.e. references to known authors, ranging from close colleagues to researchers known by name but never met), only a few characteristics could differentiate them, based on the traces left in the text by the authors. Together, they make up between half and two-thirds of citations (more in mathematics, less in biology and sociology). Thus, most citations reflect social links between authors. This opens up a new perspective for a better understanding of the role of communities in the production of scientific texts.

Citations of co-workers from the same lab tend to be repeated more often, while the trend tails off in the direction of unknown authors. We can conclude that, to some extent, the closer the cited author is to the citing author, the more they will be cited in the text.

Citations clusters – in which the author cites an indistinct set of references – tend to be homogeneous in terms of relationships. This means that the textual equivalence of citations corresponds, in part, to the relational equivalence of their authors.

Furthermore, the textual context seems, in part, to be influenced by the social relations embedded in citations. Although we were unable to identify any specific characteristics for the different social links, it is clear that citations where the social relationships are similar have textual contexts that are more similar than others.

The number of citations has become an indicator that is frequently used by many institutions to evaluate their researchers (seen, for example, in the success of the *h-index*). As this article shows, in-text references are more-or-less salient. This raises the question of whether it is really relevant to count the number of citations obtained by a researcher, without considering the role of the reference in the texts.

Moreover, this study shows that there is a correlation between the textual characteristics of the citation and the relationship between the authors. This strongly suggests that it is impossible to consider citations as objective indicators, with no social bias. Our work reveals the overlap between citations and social networks, in the sense that the social structure is partially reflected in the textual features. This suggests that inequalities between researchers regarding their social and relational capital (notably seniority or gender), can also lead to differences in the positioning of the citation in the text, and impact the citation visibility.

Last, this study highlights a diversity of citation practices across disciplines. We conclude that it is very difficult to have the same understanding of citations in such different disciplinary contexts. Although this finding is frequently reported in the literature, it is still given little attention by institutions engaged in research evaluation.

Although pioneering, our study is limited in terms of disciplinary coverage and amount of data. However, the results are promising, and confirm that citation behaviour is a very complex phenomenon that is influenced by a number of factors. In particular, the relationships between authors seems to play an important role, which cannot be neglected in further studies.

6 References

Agresti, A. (2013). *Categorical data analysis*. Wiley.

Baldi, S. (1998). Normative versus social constructivist processes in the allocation of citations: A network-analytic model. *American Sociological Review*, *63*(6), 829-846.

Bertin, M., Atanassova, I., Gingras, Y. & Larivière, V. (2016). The invariant distribution of references in scientific articles. *JASIST*, 67(1), 164-177.

Bertin, M., Atanassova, I., Lariviere, V. & Gingras, Y. (2013). The distribution of references in scientific papers: an analysis of the IMRaD structure. *Proceedings of the 14th International SSIC*, 591-603.

Bertin, M. & Atanassova, I. (2014). A study of lexical distribution in citation contexts through the IMRaD standard. *Proceedings of the Bibliometric-Enhanced Information Retrieval Workshop* at European Conference on Information Retrieval, 5-12.

Bornmann, L. & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45‐80.

Boyack, K.W. van Eck, N.J., Colavizza, G. & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis, *Journal of Informetrics*, 12(1), 59--73.

Brooks, T. A. (1985). Private acts and public objects: An investigation of citer motivations. *JASIST*, *36*(4), 223-229.

Chubin, D. E. & Moitra, S. D. (1975). Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science*, *5*(4), 423‐441.

Crane, D. (1972). *Invisible Colleges. Diffusion of knowledge in scientific communities*. Chicago: The University of Chicago Press.

Cronin, B. (1984). *The Citation Process: The Role and Significance of Citations in Scientific Communication*. London: Taylor Graham.

Cronin, B. & Shaw, D. (2002). Identity-creators and image-makers: Using citation analysis and thick description to put authors in their place. *Scientometrics*, *54*(1), 31‐49.

Ding, Y., Liu, X., Guo, C. & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics*, 7(3), 583-592.

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X. & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *JASIST, 65*(9), 1820-1833.

Glänzel, W. & Thijs, B. (2004). The influence of author self-citations on bibliometric macro indicators. *Scientometrics*, *59*(3), 281‐310.

Harwood, N. (2009). An interview-based study of the functions of citations in academic writing across two disciplines. *Journal of Pragmatics*, *41*(3), 497‑518.

Hernández-Alvarez, M. & Gomez, J.M. (2016). Survey about citation context analysis: tasks, techniques, and resources. *Natural Language Engineering,* 22(3), 327-349.

Hu, Z., Chen, C. & Liu, Z. (2013). Where are citations located in the body of scientific articles? A study of the distributions of citation locations. *Journal of Informetrics*, 7(4), 887-896.

Johnson, B. & Oppenheim, C. (2007). How socially connected are citers to those that they cite? *Journal of Documentation*, *63*(5), 609‑637.

Larivière, V., Archambault, É., Gingras, Y. & Vignola-Gagné, É. (2006). The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *JASIST*, *57*(8), 997‑1004.

Liu, J. S., Chen, H. H., Ho, M. H. C. & Li, Y. C. (2014). Citations with different levels of relevancy: Tracing the main paths of legal opinions. *JASIST*, *65*(12), 2479-2488.

MacRoberts, M. H. & MacRoberts, B. R. (1984). The negational reference: Or the art of dissembling. *Social Studies of Science*, *14*(1), 91‑94.

Milard, B. (2014). The social circles behind scientific references: Relationships between citing and cited authors in chemistry publications. *JASIST*, *65*(12), 2459‑2468.

Moravcsik, M. J. & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, *5*(1), 86‑92.

Piao, S., Ananiadou, S., Tsuruoka, Y., Sasaki, Y. & McNaught, J. (2007). Mining opinion polarity relations of citations. In *Proceedings of the International Workshop on Computational Semantics (IWCS)*, 366-371.

Ritchie, A., Robertson, S. & Teufel, S. (2008). Comparing citation contexts for information retrieval. In *Proceedings of the 17th ACM conference on Information and knowledge management*.

Schneider, J. W. & Borlund, P. (2004) Introduction to bibliometrics for construction and maintenance of thesauri: Methodical considerations. *Journal of Documentation*, 60, 524-549

Snyder, D. & Bonzi, S. (1998). Patterns of self-citation across disciplines (1980–1989). *Journal of Information Science,* 24 (6), 431–435.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Tahamtan, I., Safipour Afshar, A. & Ahamdzadeh, K. (2016). Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics*, *107*(3), 1195‑1225.

Tanguy, L., Lalleman, F., François, C. Muller, P. & Séguéla, P. (2009). RHECITAS: Citation analysis of French humanities articles. *Proceedings of the International Corpus Linguistics Conference*, Liverpool.

Teufel, S. (2010). *The structure of scientific articles: Applications to citation indexing and summarization.* University of Chicago Press.

Teufel, S., Siddharthan, A. & Tidhar, D. (2006). Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, 103-110.

Teufel, S., Siddharthan, A. & Tidhar, D. (2009). An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue,* 80-87.

Tuire, P. & Erno, L. (2001). Exploring invisible scientific communities: Studying networking relations within an educational research community. A Finnish case. *Higher Education*, *42*(4), 493‑513.

Wallace, M. L., Larivière, V. & Gingras, Y. (2012). A small world of citations? The influence of collaboration networks on citation practices. *PLoS ONE*, *7*(3), e33339.

White, H. D. (2001). Authors as citers over time. *JASIST*, *52*(2), 87‑108.

White, H. D., Wellman, B. & Nazer, N. (2004). Does citation reflect social structure?: longitudinal evidence from the "Globenet" interdisciplinary research group. *JASIST*, *55*(2), 111‑126.

Zhu, X., Turney, P., Lemire, D. & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *JASIST*, 66(2), 408-427

Zuccala, A. (2006). Modeling the invisible college. *JASIST*, *57*(2), 152‑168.

APPENDIX 1

The interview schedule.

I – Career

- How has your career developed since your PhD?
- Date the PhD was defended/ date of recruitment / other important dates
- What has been your geographical mobility?
- Evolutions, reasons for change in research themes

II – The article

Role of this article in the research career: what does this article highlight? links with previous articles?

- Collaborators:
    - o who are they?
    - o how long have they known each other?
    - o circumstances of the meeting?
    - o nature and quality of the collaboration?
- distribution of work: who did what?
- acknowledgments: who? why?
- submission:
    - o choice of journal?
    - o review process?
    - o who are the reviewers?
- the legacy of the article: conference communications, citations, etc.; what happened after the article was published?

III – References in the article

Review all of the authors listed in the references (the bibliography):

1) Can you characterise the type of relationship you have with this person whose reference is cited:
    - how do you know them? (from well-known to known by name)
    - since when?
    - what were the circumstances of the meeting?
    - what form does your discussions take?

2) What do you know about him/ her?
    - disciplinary specialty
    - geographic origin (most accurate)
    - institution, group, team

- grade (permanent researcher, doctoral student, industrialist, etc.)