Hooper, Clare J.; Peters, Isabella; Robin, Cécile

Article — Manuscript Version (Preprint)

Mapping the Topics and Intellectual Structure of Web Science

The Journal of Web Science

Suggested Citation: Hooper, Clare J.; Peters, Isabella; Robin, Cécile (2019) : Mapping the Topics and Intellectual Structure of Web Science, The Journal of Web Science, ISSN 2332-4031, L3S Research Center and Leibniz Universität Hannover, Hannover, Vol. 5, Iss. 1, pp. 1-15, https://doi.org/10.34962/xhmm-0425

Mapping the Topics and Intellectual Structure of Web Science

Clare J. Hooper
Independent Consultant
Vancouver
Canada
clare@clarehooper.net

Isabella Peters
ZBW Leibniz Information Center for Economics & Kiel University
Germany
i.peters@zbw.eu

Cécile Robin
Insight Centre for Data Analytics
NUI Galway
Ireland
cecile.robin@insight-centre.org

ABSTRACT

This paper charts the evolution of Web Science as a research field. It describes a mixed methods analysis of papers published at the ACM Web Science Conference series from 2009 to 2016, using co-citation analysis, bibliographic coupling, natural language processing, topic modelling and network visualisation techniques in order to map the intellectual structure, i.e. current topics, knowledge base and knowledge transfer, of the field of Web Science. The knowledge base of the Web Science community and the knowledge transfer from the ACM Web Science Conference series are studied, revealing major themes and key authors as a map of Web Science. In particular, the foundations of the Web Science community are revealed via co-citation analysis of authors of papers cited by ACM Web Science papers, while NLP analysis reveals topical descriptors and application contexts of Web Science. Finally, author-based bibliographic coupling of papers published at Web Science Conferences reveals authors who have been influenced by the Web Science community. In sum, this paper presents a knowledge map of the Web Science discipline visualizing topical foci, methodical roots in various disciplines, and key players in Web Science research.

KEYWORDS

Web Science; community analysis; bibliometrics; disciplines; Saffron; application contexts; mixed methods; interdisciplinarity; citation analysis; natural language processing; topic modelling.

1 INTRODUCTION

Web Science (WebSci) concerns the impact of the web and its ecosystems on society, and of society on the web. One of its key strengths is its interdisciplinarity (Hall et al., 2016), but this can also be a challenge: collaboration across disciplinary boundaries can be difficult and time consuming, acknowledged by efforts in the WebSci community to address this challenge (Hooper et al., 2014). Although there has been much discussion about the composition and representation of disciplines within WebSci, little work has addressed it. There is thus little evidence about what research foci and disciplines constitute WebSci, the true extent of its interdisciplinarity, or the topics addressed in WebSci research.

Typically, self-reflective examinations that provide answers to those questions fall under the umbrella of "Science of Science" (Fortunato et al., 2018) which provides the methods to study and map the structures of science on various levels, e.g. in terms of discipline or active scientists, thus revealing detailed insights into the intellectual basis of a discipline, its evolution over time, the community shaping it, and the specific research landscape that has been formed. These descriptive studies help build communities, defining for example who a web scientist is, and act as diagnostic tools aiding strategic management of a community or discipline. On a pragmatic level, the results of "science of science" studies help steer conference organization and align expectations with reality (e.g., by suitably guiding topics in the Call for Papers to strategically adress stakeholders).

In 2012, we made a first attempt in this regard and presented a proof-of-concept of how we can use Natural Language Processing (NLP) to understand disciplinary representation within WebSci (Hooper et al., 2012); in 2013, we built on this with an analysis of almost 500 articles and an expert survey to gain insight into links between NLP-generated terms and disciplines (Hooper et al., 2013).

The survey described in the 2013 paper asked experts to match disciplines (from past WebSci Calls for Papers) with terms (extracted from WebSci papers with NLP). We were surprised to receive unsolicited comments criticising both lists: "the [term] list seems to be very much slanted towards technology and away from anything like law, economics, sociology"; "you need to add all the [humanities] disciplines if you're going to add philosophy […] And what about art, design, media studies, gender studies?"; "There are some startling absences, e.g. business studies, art, culture […] and education."

These comments showed a discrepancy between expert perceptions of what WebSci is compared with both disciplines from WebSci Calls for Papers and core WebSci research terms. This motivated us to investigate more deeply and to detect the underlying schools of thought or "invisible colleges" (Crane, 1972) that build the basis for WebSci.

We combine several approaches to study: where the WebSci Conference community derives its knowledge (via

reference analysis); what is discussed and how knowledge is phrased (via terms from papers published by the community); and how the WebSci community's knowledge is processed (via citation analysis; Garfield, 2004).

To do so, we expanded the explorative study by rerunning the NLP analysis with more data (expanding the corpus to 2016) and applying further methods, i.e. citation and network analysis. NLP helps reveal concepts underlying current WebSci research, but WebSci papers' reference lists and citations are also rich sources that indicate knowledge flows between articles (references refer to the knowledge base that an article builds on, whereas citations reflect the impact of an article on subsequent work, i.e. the knowledge transfer).

These knowledge flows of WebSci can be exploited in our analysis/mapping of invisible colleges, interdisciplinarity and the intellectual structure of WebSci ('Intellectual structure', a term from bibliometrics, concerns how a field's structure emerges based on scholarly works cited by authors publishing in this field). By triangulation of methods from both NLP and citation analysis, we aim at overcoming potentially subjective perceptions of WebSci, as might be apparent in interviews conducted in our previous work (Hooper et al., 2013).

As such the paper takes a descriptive approach shedding light on underlying structures of WebSci and opening up to interpretation. We hypothesize that analyses will especially show that WebSci, as of today, may rather be a multidisciplinary field, leaving disciplines tackling research questions in rather isolated manner, than one of strong interdisciplinarity, in which disciplines rather work together to provide solutions for shared research problems, as envisioned by Hall, Hendler, and Staab (2016) for WebSci.

In particular, this paper addresses the following research questions:

1. What disciplines and traces of interdisciplinarity are evident in WebSci research?
2. How does WebSci link thematically with other disciplines?
    a. On what themes does WebSci research build?
    b. What areas are affected by WebSci research?
3. Who are the key people in WebSci research?
4. What does a map of WebSci reveal?
5. How has the landscape of WebSci evolved over time?

To the best of our knowledge this paper is the first analysis of its kind of WebSci. Of course, as the field progresses and expectations and reality further develop, similar studies should be carried out on a regular basis in order to follow the evolution of the field through time. We hope that such analyses support and inspire the WebSci community to further investigate and discuss its self-image and composition as well as whether its goals are met or adjustments are needed.

2 BACKGROUND AND RELATED WORK

We are especially concerned with methods that reveal knowledge flows and themes within WebSci. Thus, our work uses both co-citation analysis (Chen and Carr, 1999; Goodrum et al., 2001) which has been proven to be good for examinations of conferences (Agarwal et al., 2017) as well as journals (Haustein and Larivière, 2014) and fields (Düzyol, Taşkın and Tonta, 2010). Since our previous publication (Hooper et al., 2013), new work include an approach to visualising a domain's literature based on co-readership (Kraker et al., 2013), visualising conference-author-coupling and conference-user-coupling networks (Ni and Jiang, 2016) and mapping an interdisciplinary Network of Excellence (NoE; Sahal et al., 2013). The latter work examined where members of the NoE publish. By contrast, we focus on the themes present in WebSci publications, rather than the domains WebSci experts publish in.

We also use bibliometric methods to find traces of interdisciplinarity as well as of the intellectual structure of WebSci as appearing in articles published at WebSci conferences (Wagner et al., 2011). Since there are various publication outlets that can be addressed by an interdisciplinary audience we investigate publications stemming from the WebSci conference series, which is the main venue at which web scientists gather[1]. The Microsoft Academic Graph[2] started to compile bibliometric information on conferences too: however, these data describe top cited authors or citation venues and do not offer maps of science.

2.1 Bibliometric Mapping and Natural Language Processing

The work described in this paper is based on the assumption that research fields can be described through the use of specific keywords, as multi-word terms. This assumption relates to van Eck and Waltman's (2010) study on bibliometric mapping, where they describe how the different areas of expertise can be represented and identified thanks to clusters of keywords in a two-dimensional space.

Previous techniques addressing the domain-specific term extraction task have typically involved the use of external resources (Bordea, 2013), either as training data

[1] Another venue worth examining is the Journal of Web Science (Online ISSN: 2332-4031) that launched in 2015 and has published 15 articles till June 2018.

[2] https://www.microsoft.com/en-us/research/project/academic/articles/www-conference-analytics/

for supervised machine learning, or to build symbolic rules for non-statistical systems, like pre-built domain taxonomies (Coulter, Monarch, and Konda, 1998), or also for experts connections with sets of key phrases associated with authors specialized in a particular field, as available in some specific publication platform like Google Scholar[3], Semantic Scholar[4], or Research Gate[5]. However, such resources are often not available for the domain, or are very restrictive. In fact, there is no such resource readily accessible for the WebSci domain. More recent work from Jiang, Endong, and Jianzhong (2015) also raised this issue and proposed a domain independent technique for term extraction, leveraging the problem of human effort to create resources when they are not already available. However, their approach relies on specific common structural features derived from research papers.

Our corpus, as described below in more details, not only contains articles but also keynotes and workshops, which structures are different from research papers. Their approach would thus force us to reduce the corpus to papers only, which would result in a significant loss of information. On the contrary, the system chosen for this domain-specific term extraction work, namely Saffron (Bordea, 2013), does not restrict its scope to a type of document but can be applied to any machine readable textual documents. It was developed following an automatic method (Bordea and Buitelaar, 2010), described in the following section, which has the benefit of not relying on any external knowledge in order to extract domain-specific terms.

Word co-occurrence analysis, a content analysis technique, is then used to discover implicit relations between the extracted topical descriptors. This technique was applied to analyse the interconnections between a main field, i.e., fuzzy logic theory, and its computing techniques (Lopez-Herrera et al., 2010), a setting that is similar to our analysis of WebSci. A more recent work on co-word analysis (Wang et al., 2012) outlined several limitations related to the use of keywords and proposes a method to integrate expert knowledge into the process, requiring however a considerable amount of human intervention for the construction of domain specific thesauri.

We alleviate this challenge by completely automating the process of identifying topical descriptors and by automatically constructing a domain taxonomy (topical hierarchy), visualizable in a graph. Involving human experts for this task would have involved a colossal work, being costly in terms of both time and human expertise resource. Moreover, one can argue the ability of a human to be able to comprehend a hierarchically organized overview of the topics on such huge amount of documents, and evaluate the semantic relatedness of the terms over the whole corpus.

WebSci is an interdisciplinary field, at the crossroad of domains as diverse as Physics, Psychology, and Economics. Each domain has a different level of formality, with a varying number of natural language terms and a more or less deterministic syntax. This impacts the performance of term extraction tools, with a larger number of correct terms extracted for some domains than for others. In Zhang et al. (2008), different term extraction approaches are evaluated over two domains, a Biology corpus and a small general knowledge corpus of Wikipedia articles; term extraction performance is shown to vary depending on the domain. More recent work (Bordea et al., 2013) studies the performance of term extraction systems over three domains (Computer Science, Biomedicine, and Food and Agriculture). That work showed that Saffron, our NLP tool, produces stable results across different domains, which also motivated its reuse here.

2.2 Citation Analysis on Author Level

We use citation analysis (co-citation and bibliographic coupling) to detect knowledge flows to and from the WebSci community and to study where the WebSci community positions itself in terms of where its knowledge stems from (knowledge base) and how it is perceived by third parties (knowledge transfer). Besides taking a descriptive approach with counting citation numbers, co-citation and bibliographic coupling exploit the directed network of citations/references and publications and link publications according to the amount of shared citations/references. Co-citation analysis provides maps of scientific fields and reveals their underlying communities as well as their intellectual structure (Tonta and Düzyol, 2010). It allows for different units of analyses (i.e. levels of aggregation), e.g. journals (Culnan, 1987), authors (McCain, 1986), conferences (Ni and Jiang, 2016) or fields (Zhao and Strotmann, 2008).

In our case, two units are co-cited (Small, 1973) if they both are referenced in the same article stemming from the WebSci Conference paper corpus. This demonstrates how authors of the WebSci Conferences link units and how they utilize them for WebSci research. As such, information on co-citation reflects the knowledge base of a unit. Bibliographic coupling is the counterpart of co-citation (Kessler, 1963). Two units are bibliographically coupled if they both share at least one unit of the WebSci Conference paper corpus in their bibliography. Bibliographic coupling represents a transfer of knowledge and explains how WebSci papers are reused by other units. Hence, again, bibliographic coupling reveals how a unit influences other units and how it links them thematically.

We use the author as our unit of analysis. As described in White and Griffith (1981) and confirmed by

[3] https://scholar.google.com
[4] https://www.semanticscholar.org
[5] https://www.researchgate.net

McCain (1986), Zhao and Strotmann (2011), and Rorissa and Yuan (2011), author co-citation networks can accurately represent the intellectual structure of a field, i.e. which thought leaders the field is based on, or who is building bridges between different communities.

Following the approach of White and Griffith (1981),), we do not carry out our analyses on single documents but on sets of documents associated with a single researcher. Hence, we work with the œuvre of a researcher that reflects what the author "Strohmaier", for example, stands for in terms of topic[6]. Nerur, Rasheed, and Natarajan (2008, p. 322) conclude that "Often an author's work over a period of time tends to be characterized by thematic consistency, advocacy of a particular perspective, and cumulative contributions in answering a specific research question". Åström has shown a good correspondence between maps based on author-co-citation analysis and on co-occurrence of keywords used to describe their papers (Åström, 2002).

This author-based approach is also useful for overcoming challenges caused by sparsity of our data. The low number of papers in the citation analysis corpus and the interdisciplinary nature of the WebSci community affect the probability of receiving reasonable results on a per-paper basis (i.e. we are unlikely to detect a sufficient amount of papers that are cited and referenced more than once, given that citation distributions often obey power laws; Seglen, 1992).

The number of articles and the number of authors citing WebSci papers of the three conferences was low, so the processing of bibliographic coupling was also carried out on an author basis and for the full set of citing documents (that cite articles from each of the three conference years). Hence, author-level citation analyses allow for exploitation of denser citation distributions.

3 METHOD

We gathered two corpora of data[7], then used NLP to extract topics and conducted citation analysis, followed by graphing and visualisation techniques to make sense of the resulting graphs. We use two approaches to examine the links between papers and topics within the WebSci community: betweenness centrality of top terms from papers, and co-citation of papers. Those different approaches are complementary to one another and focus on different aspects - thus offering the chance to examine data about the WebSci community through different lenses.

3.1 Data Gathering

The NLP corpus consists of WebSci Conference proceedings from 2009 to 2016 inclusive. The precise composition of these vary year-on-year, depending on decisions by each program committee regarding inclusion of keynote talks and workshop proposals and papers. We include everything classified as part of a WebSci proceeding, giving over the 8 years a total of 778 PDF files (each file representing one paper, poster, workshop or keynote), with variation per year depending on the conference scale and whether, for example, the PC chose to include materials such as keynote talks and posters in the formal proceedings (from 2009 - 2016, respectively, we had 119, 109, 180, 45, 106, 65, 71 and 83 files: a total of 778, of which 718 were processed with our software).

The citation analysis corpus consists of publication and citation data from the bibliographic database Scopus[8]. By time of investigation, beginning of 2016, Scopus has indexed three out of eight WebSci Conferences, i.e. WebSci14, WebSci15 and WebSci16. The number of papers published at these conferences (n=198) is relatively stable with 63, 68 and 67 papers in 2014, 2015 and 2016 respectively, authored by 186, 204 and 193 authors respectively. Given the short citation window it is reasonable to report citation numbers for papers from WebSci14: 39 papers were cited 143 times. From the three years of WebSci Conferences under study 151 Scopus papers have cited at least one WebSci publication.

3.2 Natural Language Processing Method

We processed the NLP corpus with Saffron[9], a system which extracts terms and their semantic relatedness within areas of expertise in research communities. There were several benefits in choosing Saffron for this task: we did not need external knowledge, it did not require additional manual work, it runs with any machine readable corpus of text, it is domain independent, and it produces stable results in multidisciplinary fields. Further, having successfully used Saffron in our previous analysis, continuing with the same tool allows for greater consistency and comparison between the old and new analyses.

Saffron uses algorithms for domain specific term extraction and topic taxonomy construction (Bordea et al., 2013; Bordea, 2013). The system uses information extracted from unstructured documents with Natural Language Processing techniques. For a corpus analyzed, it considers the whole collection of documents as belonging to one domain which is automatically defined and delimited. This allows the identification and targeting of expressions that are specifically relevant for this area, filtering out the ones that may be used across domains or in

[6] However, the papers considered here might not be equivalent to the total œuvre of a researcher because some papers might not be co-cited/bibliographically coupled with other papers in our data set (White and Griffith, 1981).

[7] The data will be made available on Zenodo.

[8] https://www.elsevier.com/solutions/scopus

[9] http://saffron.insight-centre.org/

one precise article only. For this purpose, Saffron creates a model of the domain (or domain model), based on the corpus and defined as a vector of single words representing the most generic concepts of the area. This term extraction phase is driven by specific features and linguistic patterns, following a domain coherence approach. It first selects candidate words, placing a higher weight on nouns as they carry content meaning, focusing on single words as longer expressions are too specific for the scope of a domain model as defined above, and being distributed in more than a quarter of the corpus (see Bordea et al., 2013 for more). A filtering phase then follows, driven by the assumption that generic domain words are often mentioned alongside many more specific domain words. Once the main high-level concepts of the domain have been collected, i.e. the domain model, the system extracts more specific terms (multi-word noun phrases) belonging to the domain. The domain model is thus used as a base to evaluate the coherence of the candidate terms within the domain using semantic similarity by Pointwise Mutual Information calculation.

The analysis was run on 718 files, composed of 20 workshops and a mix of 698 papers, posters and keynotes, from which Saffron yielded 9095 potential topics. Such an amount of terms is neither intelligible nor interpretable. Therefore, after ranking the topics based on a combination of statistical measures described in details in Bordea (2013), we focused on the most meaningful ones, working with the 500 top ranked ones in order to keep a high level of quality for construction of the taxonomy. Out of the 718 documents, 650 contained at least one of those top 500 topics, and on average 15 different topics were found in each document.

To create the pruned graph which represents the taxonomy, we adopted a similarity measure known as the association strength (or proximity index or else probabilistic infinity index), based on studies on bibliometric mapping techniques (Van Eck and Waltman, 2010). Van Eck and Waltman (2009) showed the interest in using this similarity measure over other well-known measures, such as the cosine and Jaccard indexes. They demonstrated that the latter, belonging to the category of set-theoretic similarity measures, does not properly correct for the size effect as opposed to probabilistic similarity measures, which the association strength belongs to. As a result, the set-theoretic measures do not properly normalize co-occurrence data. Therefore, in this study we are using the association strength which is defined as the measures of the strength of relationships between two research terms:

$$I_{ij} = D_{ij} / (D_i D_j) \quad (1)$$

where D_i is number of articles that mention the term T_i in our corpus, D_j is number of articles that mention the term T_j, and D_{ij} is the number of documents in which both terms appear. Edges are added in the terms graph for all the pairs that appear together in at least three documents. Saffron uses a generality measure to direct edges from generic concepts to more specific ones. This results in a dense, noisy directed graph that is further trimmed using an optimal branching algorithm which was successfully applied for the construction of domain taxonomies in Navigli et al. (2011). This yields a tree structure where the root is the most generic term and the leaves the most specific ones.

We used a network graph tool, Gephi[10], to build the graph displaying the links between terms: the nodes are the extracted terms and the edges are their links with each other. This let us identify clusters of closely related terms. We used the Force Atlas 2 algorithm (Jacomy, 2011) to layout the graph with the following parameters: Scaling: 2.0; Edge weight influence: 0.0. We used betweenness centrality to weight node importance: this measures the fraction of shortest paths going through a node (a high value shows that a node plays an important bridging role in the network; Barthélémy, 2004). Finally, we ran the Louvain method (Blondel et al., 2008) with resolution 10 to detect an interpretable set of communities. The resolution of 10 was chosen because this yielded communities of a sensible granularity: more than one meaningless 'web science' community, but fewer than 100 disparate communities that each relate to very specific terms or techniques.

We interpreted detected communities as application contexts concerning topics ranging from technologies (e.g. machine learning) to disciplines (e.g. social science) and topic areas (e.g. open government).

3.3 Citation Analysis Method

We worked with VOSviewer 1.6.5 (Van Eck and Waltman, 2010) to process citation data and construct citation networks. To analyse the knowledge base of the WebSci Conferences from 2014 to 2016, we carried out a co-citation analysis of the unique authors referenced in articles published at those conferences. The knowledge transfer from authors mentioned in papers published at those conferences to 151 citing documents was analysed by using bibliographic coupling. VOSviewer was used to visualize the network of co-cited authors for the three years together and for each year separately. All cited authors were included, resulting in a co-citation-network of 5,277 authors. The author network resulting through bibliographic coupling of 151 citing documents consisted of 453 authors.

Before running co-citation analyses and bibliographic coupling, automatically extracted author names from VOSviewer were manually checked, corrected (e.g., 'commission, e.' became 'european commission'), and merged to account for different citation styles used in the

[10] https://gephi.org/

papers of the citation analysis corpus, (e.g., 'adamic, l.' and 'adamic, l.a.' were merged to 'adamic, l.a.'). This was carried out by one author (IP), comparing extracted author names with names indexed in Scopus. A conservative approach was taken: if names were too ambiguous they were not merged.

We used VOSviewer's association strength algorithm for clustering and network layout (Van Eck and Waltman, 2009). To optimize the clustering and visualizations, VOSviewer's standard settings were kept as provided for co-citation analyses and bibliographic coupling. To collect representative terms that describe the fields the authors are working and publishing on (their œuvre), we searched the authors' Google Scholar Citation Profiles and personal webpages for keywords they use themselves to characterize their research (even if they may not be the most accurate representation of the authors' œuvres or might be incomplete).

4. RESULTS

In the following we present and briefly discuss the results of the NLP and citations analyses. In Section 5 we will summarize the results and reflect on them in view of our research questions.

4.1 Natural Language Processing Results and Discussion

The NLP outputs let us visualise the extracted terms as well as examine top rated terms and application contexts. Figure 1 shows a visualisation of the extracted terms, where larger nodes and label fonts indicate terms with higher betweenness centrality and colours indicate detected communities. Table 1 lists terms with a high betweenness centrality, and includes for reference the highly rated terms from our previous work in 2013 (Hooper et al., 2013).

In 2013, as now, we generated a list of the top 20 terms as ranked by betweenness centrality. Perhaps surprisingly, only 7 concepts are in common (although some concepts take multiple positions in one or both lists, e.g. "social network", "social networking", "social networking site"). Common concepts are: linked data / linked data principle; semantic web; social science; social network / social networking / social network site; social media / social medium; information retrieval; search engine. The terms include technologies (linked data, semantic web), disciplines (social science) and core WebSci topics (social networks and social media): this is all to be expected. Some terms only appeared in the 2013 list. These are: learning network, web page, personal learning environment, social interaction, mobile device, future research, internet user, uniform resource identifier, web science research, user interface, web community, web application, linked data principle.

Table 1: The 20 terms from 2017 with highest betweenness centrality (b.c.), and for comparison the top terms from our 2013 analysis (note b.c. values are not directly comparable as they are sourced from separate graphs; Hooper et al., 2013).

2017 term (b.c.)	2013 term (b.c.)
open data (177)	semantic web (758)
science research (150)	social media (590)
open government (144)	information retrieval (504)
linked data (140)	social networking site (495)
public sector (132)	social science (456)
government data (120)	search engine (454)
semantic web (115)	social networking (434)
social science (112)	learning network (360)
social network (105)	web page (304)
data management (105)	personal learning environment (297)
preferential attachment (96)	social interaction (282)
social medium (86)	mobile device (270)
machine learning (81)	future research (260)
training centre (80)	internet user (258)
random graph (75)	uniform resource identifier (246)
information retrieval (70)	web science research (235)
computer science (61)	user interface (235)
time series (52)	web community (235)
data collection (51)	web application (234)
search engine (50)	linked data principle(231)

We believe the learning terms arise from proceedings of the 2011 Personal Learning Environments conference, which was included in our previous corpus due to their presence on journal.webscience.org.

These materials were excluded in our 2017 corpus as less relevant to WebSci, so the absence of learning terms is expected. In 2013, the terms "web page", "future research" and "internet user" were associated with disparate contexts (much like "training centre" in 2017), so their disappearance in the context of a bigger, potentially more cohesive corpus is logical. We speculate that, if we were to repeat today's analysis in 2021 with the next four years of WebSci conference data, the term "training centre" would be very likely to disappear. "Social interaction" is another 2013 term now absent. We identified it as relevant across many disciplines and reflecting the ethos of WebSci. Should we, then, be concerned about its absence -- and the absence of the comparable terms "web science research" and "web community" in 2017?

The continued presence of key WebSci application contexts suggests not, but the loss of "social interaction" suggests a shift in focus towards other topics. Other top 20

terms from 2013 which dropped to a lower ranking are "mobile device" (unclear in meaning, although associated with many disciplines) as well as terms that we argue are oriented around computer science: "uniform resource identifier", "user interface", "web application". Finally, "linked data principle" disappeared, but with "linked data" the 4th strongest term in 2017, we attribute this to changes in wording, not focus.

The terms "open data", "open government", "public sector", "government sector" and "data management" are new to the top 20 list in 2017. These terms appear in two different application contexts, which we believe are closely related. Their emergence makes sense in the context of the rise of open data. Other new top 20 terms are "preferential attachment", "random graph" and "time series", part of the graph theory context and suggesting an increased focus on graph theory. "Machine learning" and "computer science" clearly refer to computer science work, while "science research" and "data collection" are more general terms. The other new term, "training centre", is part of a small community including the term "web science doctoral training centre": we presume this arises from papers acknowledging the support of this centre.

Figure 1 shows communities revealed by the community detection algorithm, which we can examine and interpret as WebSci application contexts. Each community has a subset of terms, which can be ranked with betweenness centrality. Table 2 details the nine biggest communities (we chose to convey the most interesting nine communities in this analysis for reasons of brevity), including for each its most highly ranked five terms, additional terms of interest, number of nodes (giving an idea of scale) and colour and position in Figure 1. The largest communities are shown first.

We distinguish between "communities", detected algorithmically, and "application contexts", where we have made sense of multiple communities and their links to discern their meaning in WebSci. We note that although we describe the application contexts here as separate entities, relationships between them exist: for example, information retrieval and machine learning clearly have links with social research. We believe, however, that the links *within* an application context are stronger than those *between* separate application contexts and are thus not always connected with a link in the graph.

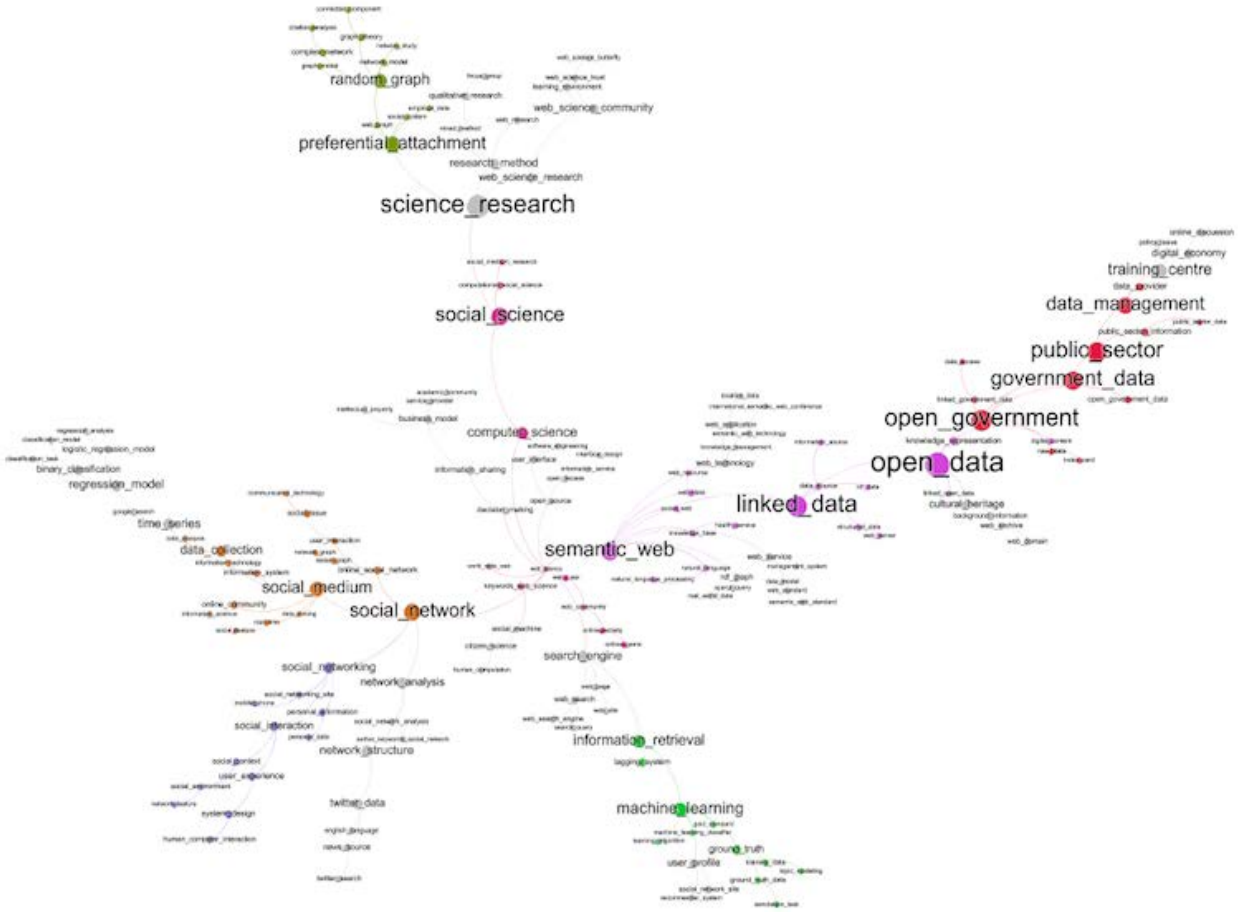

Figure 1: Visualisation of the extracted terms.

Table 2: Detected Web Science 'communities' (underscores in terms replaced with whitespace for readability).

Root node	Top 5 terms	Other terms of interest	Number nodes	Colour and position in Fig. 1
open data	linked data; semantic web; knowledge representation; web service; data source	knowledge base, information space, web ontology language, rdf triple, semantic network	41	Purple, mid right
machine learning	information retrieval; ground truth; tagging system; training data; ground truth data	classifiers, learning algorithms, annotation, reputation, linguistic features, tagging	26	Lime green, bottom right
social network	social medium; data collection; online social network; online community; information system	social graph, communication technology, content analysis, sentiment analysis, health information	26	Orange, mid left
open government	public sector; government data; data management; data provider; public sector information	linked government data, public sector data, public service, data market, service delivery	26	Red, far right
social networking	social interaction; user experience; system design; social context; personal information	human computer interaction, socio technical system, social capital, location based social network, online shopping	25	Blue, bottom left
social science	computer science, keywords web science, web community, computational social science, online activity	web science, case study, multiplayer online game, internet auction, educational technology, political science, actor network theory, micro blogging service	25	Pink, upper middle
preferential attachment	random graph; complex network; network model; graph theory; connected component	empirical data, scale free network, network property, network statistic	25	Dark Green, top of middle
network structure	network analysis; twitter data; news source; twitter search; english language	online network, internet access, health care, information flow	22	Grey, bottom centre
science research	web science community; web science research; web science trust; learning environment; web science butterfly	web science conference; web science curriculum; social science research; learning process; educational resource	14	Grey (again), upper middle

The appearance and centrality of the *open data community* is no surprise, with open linked data a central WebSci construct. Open data can be seen as both a technology and a social movement (especially in the context of government and public sector data), and has taken off in recent years. The fourth largest community, *open government*, is linked to *open data* physically on the graph but also linked via its terms. We can view linked open data as an application context of WebSci, with government and public sector data a topic within that context.

The *machine learning community* centres on methods, with terms such as "ground truth", "training data" and "classifiers". Again, it is unsurprising to see this in the context of WebSci, as these computer science methods are relevant for WebSci: machine learning is an application context for WebSci.

The *social network community* can be reasonably assumed to concern social network analysis across the breadth of WebSci, with reference to "social issue", "online community" and "social feature". *Social networking* is a community that, although identified apart from the *social network community*, appears to be part of it: it continues the thread with topics such as "social interaction", "social context" and "personal information". Similarly, the *network structure* community, connected to the *social network* one

and with terms about social network analysis, can be considered part of the overall context of social networks. The network structure context, also connected to the social network context, refers to methods such as "network analysis", and data sources ("twitter data", "news source"); we can also see this as an elaboration of the social network context.

The social science context captures several WebSci disciplines and crosses multiple topics (e.g. "educational technology", "political science" and "multiplayer online game"); it also includes reference to at least one social science approach ("actor network theory"). It is heartening to see that social science and computer science, fundamentally different disciplines that are foundational in WebSci, enjoy a direct link. This is most probably due to the popularity gained by "computational social science" over recent years. The science research community contains methodological terms such as "web science research", "mixed method" and "qualitative research" and is linked to the social science community. Together they could be seen as a "web science" application context, talking about WebSci disciplines and tools.

The preferential attachment community contains references to graph theory, including "graph model", "network model" and "web graph"; this context concerns graph theory.

We do not analyse leaf nodes (individual terms that were disconnected from other parts of the graph) and smaller communities. An example is the *web technology community*, connected to (and, we argue, part of) the semantic web context, and containing terms such as "knowledge management" and "semantic web technology". We believe these smaller contexts (small collections of nodes connected to larger contexts, rendered in grey) fall within the greater contexts with which they are connected and hence do not analyse them in turn.

The application contexts are therefore: "linked open data"; "machine learning"; "social networks"; "web science"; "graph theory". In 2013, we found four application contexts: "information retrieval"; "personalised learning/elearning"; "semantic web"; "social networking".

"Social networking" appears in 2013 and now. We note that "information retrieval" (the 2013 context) was a central term in the new "machine learning" context (and in 2013 had associated terms such as "sentiment analysis" and "knowledge management"): we may consider these two contexts synonymous. The old "semantic web" context now underlies the "linked open data" context, while the "personalised learning/elearning" context has disappeared, consistent with the 2017 corpus not including learning-specific materials. This leaves two new contexts, "graph theory" and "web science". The rise of "graph theory", again, is consistent with what we saw in the top 20 terms. The rise of a "Web Science"-specific context suggests consolidation within the community.

4.2 Citation Analysis and Discussion

In the following we present and briefly discuss the results of the author citation analyses of three Web Sci conferences. Analyses were performed for all years combined and separately for each year.

4.2.1 Citation Analysis of Authors Mentioned in WebSci Conferences 2014-2016

Figure 2a shows the results of the author-based co-citation analysis. A 'density view' is used, meaning the "larger the number of items in the neighborhood of a point and the higher the weights of the neighboring items, the closer the color of the point is to red. Conversely, the smaller the number of items in the neighborhood of a point and the lower the weights of the neighboring items, the closer the color of the point is to blue" (Van Eck and Waltman, 2016, p. 7). The density view also reveals information about how often an author has been cited (i.e. size of author name) and about the strength of the co-citation relationship of two authors. The closer two authors are located to each other, the stronger their relatedness in terms of co-citation.

The combined co-citation analysis of three WebSci Conferences resulted in 52 clusters that include 1 to 227 authors. The largest network of connected authors (giant component) consisted of 5,233 authors.

The network displayed in Figure 2a forms the knowledge base of WebSci and reveals a strong core of authors concerned with complex networks and social media and data mining (Jure Leskovec, 31 citations), computational social science (Markus Strohmaier, 13 citations), web data management (Gerhard Weikum, 21 citations), WebSci theory (Wendy Hall, 26 citations), and sociological aspects of the Web (danah boyd, 29 citations). Left from the center we also see the social machine and citizen science cluster around David De Roure (18 citations) and Chris J. Lintott (11 citations). Islands are formed by Sophie Stalla-Bourdillon (3 citations; not shown due to visibility reasons) and the legal aspects concerned with the Web, Ramesh Jain (not shown) and research on heterogeneous web data streams, Peter H. Kahn (3 citations; not shown) and psychological aspects of human-technology interaction and socio-technical systems, Bogdan State with the combination of demographic studies and web data (7 citations), and Philipp A. Schrodt (7 itations) combining the web with political concepts which is also done by David Chandler (5 citations; not shown).

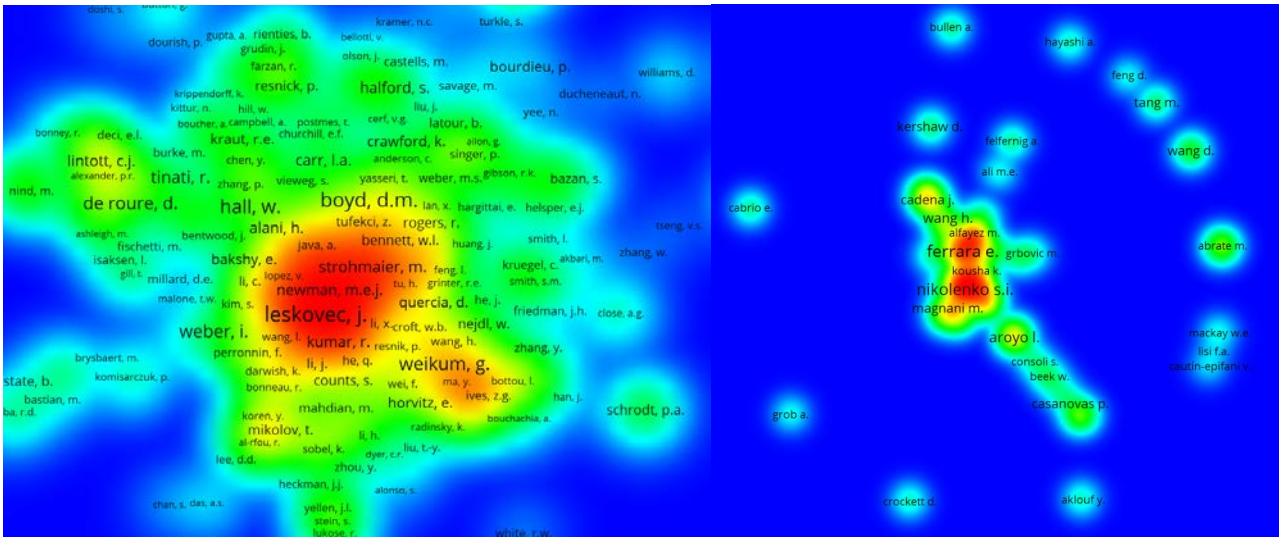

Figure 2: a) Left: Co-citation network of authors cited in articles published at the Web Science Conferences 2014-2016 (n= 5,277 cited authors); b) Right: Bibliographic coupling network of authors citing articles published at the Web Science Conferences 2014-2016 (n= 453 citing authors).

The results of the bibliographic coupling of authors citing papers from the WebSci Conferences 2014-2016 are presented in Figure 2b. The 453 authors were divided into 34 clusters each containing 1 to 34 authors. Although the giant component consisted of 388 authors we see a fragmented network with many isolated nodes (only partly shown). This reveals that reception of WebSci Conference articles takes place in different communities that are not necessarily connected to each other. The topics which are informed by WebSci research are, for example, web technologies and information visualization as represented by Matteo Abrate and Di Wang, linked data (Elena Cabrio) or studies of human behaviour (Andrew Bullen). On the other hand, we see a strong core related to, broadly speaking, data science, around Sergey I. Nikolenko (cites 5 WebSci Conference papers in articles) who works in machine learning and internet science, and Emilio Ferrara (7 citing articles) who stands for network science and computational social science. Lora Aroyo (4 citing articles) and her work on crowdsourcing, social web data, personalization and human computer interaction serves as bridge to Pompeu Casanovas (2 citing articles) who applies semantic web concepts to the law.

4.2.2 Knowledge Base of WebSci Conference 2014

The co-citation network for WebSci14 is displayed in Figure 3. The similarity processing grouped 1,734 authors in 38 clusters, which include between 1 and 116 authors. The largest cluster of connected authors consists of 1,647 authors. The co-citation analysis on the basis of authors resulted in a dense core network with Meeyoung Cha, Fabrizio Benevenuto, Krishna P. Gummadi, Filippo F. Menczer, and Jon M. Kleinberg cited most often by Websci14 authors, with 16, 13, 13, 12 and 11 citations respectively.

The area around Meeyoung Cha, (16 citations), including Krishna P. Gummadi,, (13 citations), Florent Perronnin (4 citations), and Fabricio Benevenuto (13 citations), reflects the topics of social computing, computational social science, and artificial intelligence. Interestingly, in this cluster the links between data-driven science and social science are strongly visible via danah boyd, Barry Wellmann, and Manuel Castells who mainly publish in sociology and philosophy. This area is clearly separate from the research conducted by Filippo F. Menczer and colleagues, who study complex networks with computational methods only.

The knowledge base of the WebSci14 is also characterized by fragmentation. There are several fields of study that are only loosely connected, if at all, to the core authors and topics. On the left, we see a cluster of authors that study the Web and its effect on other areas: Carl T. Bergstrom (2 citations) is concerned with the philosophy of science and scientometrics, Scott A. Hale (5 citations) studies knowledge sharing and language on the Web, Niels Brügger (2 citations) researches web history, and Ethan Zuckerman (3 citations) studies citizen media and journalism.

More distinctive islands in the co-citation network are formed by Lara Schibelsky Godoy Piccolo (3 citations), who is specialized in human-computer interaction and user engagement research, and Philipp A. Schrodt (7 citations), who works in political science and international relations. Both are examples of authors that have not been co-cited

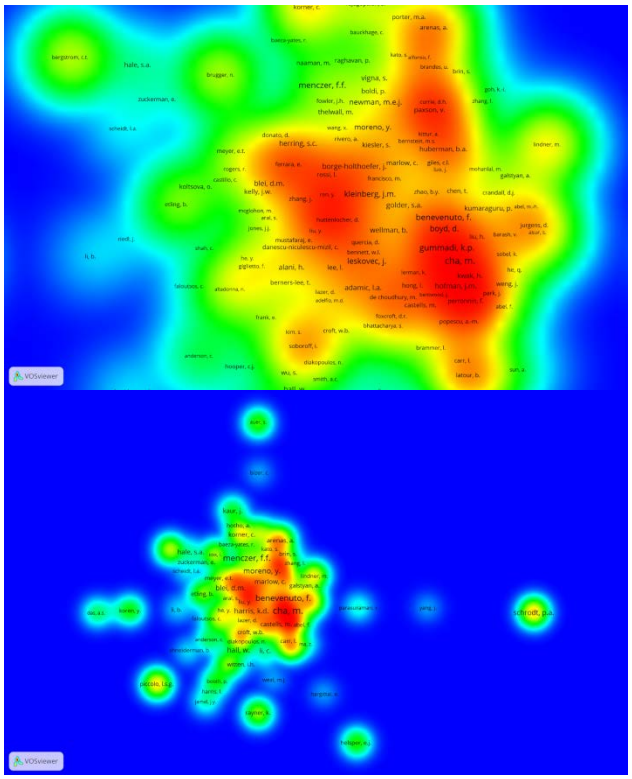

Figure 3. Co-citation network (density view) of authors cited in articles published at the Web Science Conference 2014 (n= 1,734 cited authors). a) Top: detailed view; b) bottom: full view.

with any authors of the core network, meaning that WebSci14 authors have not explicitly created a link between Piccolo and Schrodt's topics and the core network topics: no knowledge flow has taken place here.

4.2.3 Knowledge Base of WebSci Conference 2015

The co-citation network of the authors cited in the publications of the WebSci15 is shown in Figure 4. The associations algorithm identified 33 clusters, including 1 – 116 authors. The largest network of connected authors consists of 2,117 authors. Four authors influenced WebSci15 authors strongly: David De Roure, the WebSci15 program chair and Tim Berners-Lee each received 13 citations, danah boyd, the WebSci12 keynote speaker, was cited ten times, and Gerhard Weikum was cited eight papers.

We can highlight several areas of importance. Starting on top we see the community around David De Roure, concerned with social machines, citizen science, and the web of data. Close by we find Tim Berner-Lee and authors working on the theory and practice of the world wide web. This community can be considered a topical bridge between the De Roure-cluster focused on human-web-interaction and the complex network cluster represented by Jure Leskovec and Mark E. J. Newman who study properties of networks on a large scale.

The area around Gerhard Weikum and Christian Bizer involves web data management, data mining and integration, and linked data. Further right are connections to natural language processing, text mining, and machine learning (represented by Marco Pennacchiotti). The complex network-cluster and the web data management-cluster are also the densest areas of the network, reflecting strong relations between these authors. Further to the left, another topical community has formed around danah boyd: the authors associated with this cluster can be labelled 'Facebook researchers' studying this particular social network in terms of youth cultures (danah boyd), computer-mediated communication and online dating (Nicole B. Ellison), media psychology and emotional contagion (Jeffrey T. Hancock).

The co-citation analysis also revealed some areas that are rather loosely connected to the dense core topics. The cluster around Susan Halford (5 citations) connects web research with concepts from sociology and philosophy, prominently reflected by Michel Foucault. Chandler and Jirotka appear rather isolated in terms of co-citations. They mark distinct fields in web science with David Chandler (cited five times) being concerned with political concepts and Marina Jirotka (three citations) studying construction of human-computer interfaces, human-centred computing, and computer ethics. No connection to the core area of the WebSci15 knowledge base has been detected, for example, for Ramesh Jain (three citations) and colleagues researching heterogeneous web data streams.

4.2.4 Knowledge Base of WebSci Conference 2016

The co-citation analysis reveals a less fragmented landscape of the WebSci16 knowledge base than from the years before (Figure 5). It has 33 clusters consisting of 5 - 116. The dense area spans from Ingmar Weber (16 citations) to Jure Leskovec (14 citations) and represents the fields of computational social science, social media and data mining, and complex networks. Another dense cluster has developed around Gerhard Weikum (11 citations), which is concerned with web data management, web retrieval and crowdsourcing.

The third strong area is represented by authors that take a sociological or HCI-approach to WebSci and are especially interested in social media use, youth culture, and methodological questions in web research, for example Axel Bruns (10 citations), danah boyd (9 citations), and Katrin Weller (6 citations).

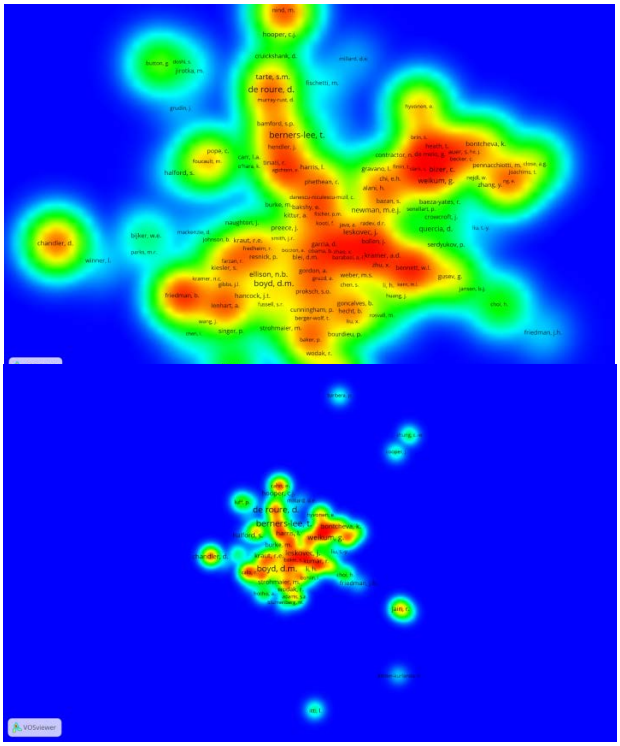

Figure 4. Co-citation network (density view) of authors cited in articles published at the Web Science Conference 2015 (n= 2,153 cited authors). a) Top: detailed view; b) bottom: full view.

Jenny Preece (4 citations) with her work on gamification, motivation, and citizen science and Rachel K. Gibson (2 citations) who studies politics organisations' use of new media, serve as topical bridges to philosopher/ social scientist Pierre Bourdieu (7 citations) and his work on power relations and different forms of capital humans can build. This area is expanded by Nicolas Ducheneaut (3 citations) and research on video games, online presence, computer supported cooperative work, and hostility on the Web.

The other area linked to web usage is represented by Kevin Crowston (4 citations) who studies open source software. Emilio Zagheni (6 citations) stands for specialized research in complex web systems, e.g. investigating digital traces for demographic research and social statistics. Isolated clusters consist of semantic web research (Harith Alani, 2 citations), legal aspects of information technology and the Web (Sophie Stalla-Bourdillon, 3 citations), and methods for empirical social science (Peter Atteslander, 1 citation).

5. Mapping Web Science

In the next section we summarize and explain the main findings of our analyses and lay out the intellectual structure of WebSci as well as our results on the interdisciplinarity of the field. We also discuss the limitations of our approach.

5.1 The Intellectual Structure of Web Science

The NLP analysis revealed the top 20 terms from the WebSci conference series, including technologies, disciplines and core WebSci topics; some concepts remain in common from the 2013 analysis ("social networking", "linked data", "semantic web", "social science", "social media", "information retrieval" and "search engine"), while some disappeared due to: exclusion of an e-learning corpus; disparate contexts disappearing; a slight drift from some computer science terms. New terms concern the rise of open data and an increased focus on graph theory.

The five application contexts revealed by NLP analysis are: "linked open data"; "machine learning"; "social networks"; "web science"; "graph theory". The "social networks" and "machine learning" contexts are broadly constant over time, and we believe the 2013 "semantic web" context underlies the "linked open data" context. New contexts are "graph theory" and "web science". The interesting results obtained by comparing the 2013 and 2016 corpora motivates us for future work to go further in the analysis and do study on a year-to-year basis.

All citation networks reveal strong connections between the authors that build the knowledge base of and are affected by WebSci Conferences. The giant component from 198 articles over the three years studied consisted of 5,233 authors, showing that authors publishing at WebSci Conferences share a high amount of authors they cite in their works. This indicates a tendency of self-referencing in the WebSci community (as represented by conference papers). Moreover, the knowledge base of WebSci is well-represented by the five most often cited authors over 198 papers published at the three WebSci Conferences: Jure Leskovec received 31 citations, danah boyd was mentioned 29 times, Wendy Hall was cited 26 times, and Jon M. Kleinberg and Meeyong Cha got 24 citations each. Those names translate, roughly, to the main topics studied by WebSci Conference contributors: Complex networks, social behaviour on the web, foundations and theory of the web, network theory and topology, social computing and (social) data science.

The three years of investigation also revealed that the core themes (as represented by authors), such as "complex networks" and "web theories", remain relatively stable which was confirmed for seven conferences hosted by ACM SIGWeb (Agarwal et al., 2017). We can also witness

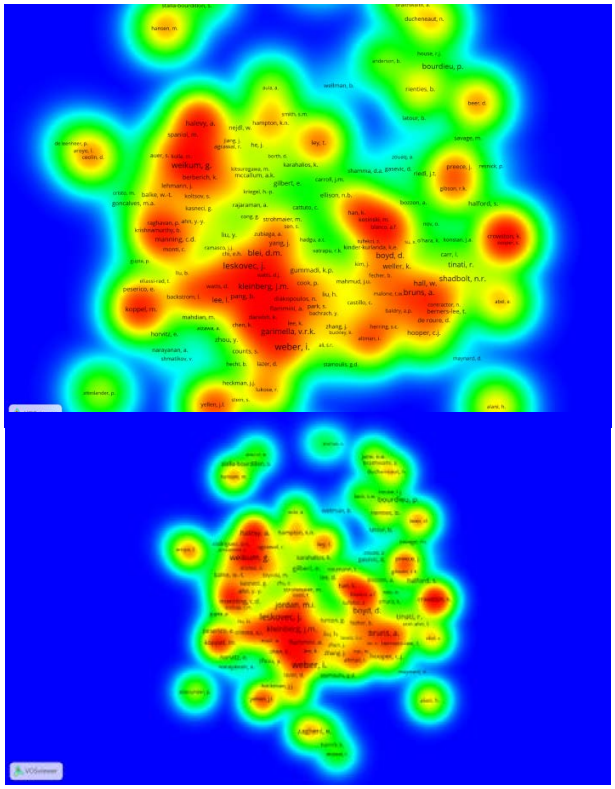

Figure 5. Co-citation network (density view) of authors cited in articles published at the Web Science Conference 2016 (n= 2,399 cited authors). a) Top: detailed view; b) bottom: full view.

changes in topics of the conference, such as gamification in 2016 or human-computer interaction that became more central over the course of the analysed years. However, the map of WebSci may be influenced by competing conferences and publication venues that are more specialized and that may affect publication behaviour of authors. Even though Agarwal et al. (2017) argue that some prolific authors of SIGWeb conferences publish at multiple conferences, our citation analysis showed that semantic web research, for example, is not too popular at the WebSci Conferences. This could also be due to the change in organisation of the WebSci Conferences: after co-location of WebSci10 with the WWW Conference and WebSci11 with Hypertext, the WebSci Conferences have been standalone events.

5.2 Interdisciplinarity

Of particular relevance in the NLP analysis is the emergence of the application context named "social science", which despite its name includes discipline names including "social science", "computer science", "political science" and "computational social science". The close linking of these terms in the graph is a positive indicator of interdisciplinarity in WebSci.

The citation analyses showed that, although there are some separate clusters concerned with specific topics and authors (Philipp A. Schrodt in Figure 2a) the networks are dense and grouped around a notable center reflecting heavy mutual citation in WebSci. Knowledge transfer from WebSci Conferences to other areas concerned with the web (e.g., human computer interaction) also takes place. This is further evidence for the interdisciplinary nature of the WebSci community.

But, interdisciplinary research that reciprocally informs each discipline is very much focused on computer science, network science, and sociology in both, knowledge base and knowledge transfer. Other disciplines, such as law or philosophy, although being recognized and discussed in the WebSci community, form islands of authors without links to the core areas and authors of WebSci. Here, there is still room for mutual learning and exchange of disciplinary concepts and methods to enhance knowledge building in WebSci and increase interdisciplinarity, as was always intended (Hall et al., 2016). Of course, the WebSci Conference series is a good venue to bring together communities and disciplines, and our analyses confirm the success of these exchanges.

However, interdisciplinary endeavours need strategic planning which also accounts for disciplinary differences, for example regarding publication and citation behaviour. WebSci16 is a good example for strong interrelations in the WebSci community, only little fragmentation took place. We assume that this is because extended abstracts, besides short papers and long papers, were accepted as submissions for the first time since 2013 and that this incentivized social scientists and disciplines with a different scientific reward system to participate.

5.3 Limitations

WebSci is a very young field, and as such the eccentricities of specific conferences (e.g. the WebSci 2011 Personal Learning Environments workshop) can strongly affect our results. We have highlighted when our results are thus affected. As the field matures and its dataset grows these effects will diminish in subsequent analyses.

We note two dataset limitations. Firstly, the use of WebSci Conference proceedings in the NLP corpus led to a slight inconsistency in document type. For example, the WebSci proceedings from 2013 and 2016 included workshop papers, but other years did not. Ideally we would have included all workshop papers, on the basis that workshops broaden the scope of the conference and are potentially venues where particularly interdisciplinary

exchanges may occur; unfortunately, archives from past conferences did not allow this. Secondly, the citation analysis dataset from Scopus includes only three years of WebSci Conferences (2014-2016). Hence, the knowledge transfer study was only possible for the 2014 conference because of the low number of citations for articles published at the 2015 and 2016 conferences. The advantage of Scopus is, however, that it indexes author surnames and initials, making it an unambiguous and well-proven approach that is a strength for author-based citation analysis.

Moreover, limitations concern the methods used in our study. Besides the positive feedback loop increasing the likelihood that highly cited articles are cited again (Merton, 1968), citation analysis in general suffers from uncertainty regarding the motivation of underlying citations and its effect on the developing citation article network (e.g., authors tend to cite close colleagues more often; Cole and Cole, 1973). Authors may also work on several topics at the same time so that the œuvre represented in this limited WebSci set may only acknowledge some of the authors's research foci. Citation cartels and extensive self-citation distort the network the most. Our results revealed that both strong cores and islands in citations networks might be symptoms of these practices. But, we have not excluded self-citations from our analyses since we were interested in the actual knowledge base and transfer that also comprises authors who strongly build their research on their own work.

We also did not control for co-authorship. There is a tendency towards higher numbers of co-authors in the sciences (Wuchty et al., 2007) and WebSci Conferences have the highest number of co-authors among seven web research-related conferences (Agarwal et al., 2017). However, this effect is highly discipline-specific. If using authors (and their publications; i.e. their œuvre) as proxies for WebSci themes, we can assume that this tendency dilutes the results of the citation analysis as author profiles may lose their discriminatory power – but this is currently not confirmed.

Finally, we note that this discussion only concerns 8 years of Web Science. Most papers experience a peak in citations 2-5 years after publication (depending on the discipline; Brody et al., 2006), and so further insights are likely to arise as time passes and citation networks further consolidate. Such a limitation is inherent to any temporal analysis of a field, and we recommend that this kind of analysis is repeated regularly to gain improved insight.

6. CONCLUSIONS

We have applied techniques from NLP and network analysis in conjunction with citation analysis on a significantly large corpus of WebSci materials (698 articles and workshops), as well as discussing how these methods complement one another. We have also carried out a domain analysis of the Web Science field to see how its intellectual structure has changed over time. By combining several approaches, examining both terms and co-citation networks, we have been able to examine both topical trends and knowledge flows in WebSci. By using mixed methods, a principle at the heart of WebSci, we have gained qualitative and quantitative insights into a disciplinary representation within WebSci, key themes and invisible colleges.

WebSci as a field is very young: there is only 8 years of data that could be analysed in this paper. We argue that WebSci's youth means such analysis is particularly important, helping both define and refine the community. Such analyses go beyond anecdotal evidence about the "look and feel" of the WebSci Conference series but they can act as diagnostic tools; and as such can inform decision-makers and drive strategic decisions based on hidden structures revealed by real data, for example, in terms of outreach to disciplines within WebSci (this may take the form of how we write CFPs, host workshops, invite keynote speakers, and co-locate events), our understanding and development of WebSci curricula, insight into changes in WebSci over time, allowing evaluation of the impact of past decisions and refining future activities accordingly. In the hands of Web Science experts, this information can help them maximise the efficacy and impact of WebSci or guide WebSci towards more interdisciplinarity, for example.

We hope that this analysis of WebSci at an early stage will also be relevant for future work as researchers follow its evolution through the years. Finally, we also hope these insights will be of pragmatic use, e.g. for the organisers of future conferences.

In addition to such contributions, this work provides insights in response to our questions from Section 1 that can be found in this paper as follows:

- Evidence demonstrating some interdisciplinarity in WebSci research: Section 5.2.
- Major themes in current WebSci research: Section 4.1.
- Themes on which WebSci research builds: Section 5.1.
- Areas affected by WebSci research: Section 5.1.
- Key people in WebSci research: Section 4.2.
- A map of WebSci's intellectual structure and its evolution over time: Sections 4.1. and 4.2.

To sum up, NLP showed the top 20 WebSci topics (and the shift in these in four years) as well as application contexts and the change in these in four years (notably, the

rise of the open data movement, close connections between different disciplines, and the consolidation of WebSci as a context of its own). Citation analysis has shown that the WebSci community is well connected and that underlying structures, such as invisible colleges, have formed around particular authors, such as Jure Leskovec and Wendy Hall, that also represent typical themes of WebSci such as network science, data science, data mining.

The comparison of the knowledge flows over time in WebSci revealed that the knowledge base is getting denser and that there are less isolated networks in the WebSci Conference 2016, for example. This can be an effect of either a consolidation of the field or of the change in the conference organization that increased its attractiveness for, amongst others, social scientists. However, the islands that remain are interesting (often concerned with law, political science), especially in terms of reflection on the multi- or interdisciplinarity of WebSci and of introducing measures that encourage publication of interdisciplinary work at WebSci Conferences (if this is a strategic demand of the WebSci community at all).

The two analyses give similar overarching results with respect to consolidation of the community and the focus on network and data science topics. One aspect of possible concern here is the lack of data suggesting centrality or connectedness of other disciplines, such as philosophy, law or psychology, have found evidence of interdisciplinary links that may be strengthening (Section 5.2).

To our knowledge, this is the first time NLP and citation analysis have been used together for analysing WebSci. They complement each other in certain ways: firstly, while NLP analysis takes a snapshot of the time span in which the dataset is based and the current WebSci themes, citation analysis allows an examination of the time before the dataset (the knowledge base of WebSci) and the time afterwards (the knowledge transfer from WebSci).

Another way the two methods work together concerns bias. Participants of our 2013 survey were concerned the NLP-generated terms were biased towards technology: it is still unclear whether this apparent bias is in the nature of NLP or an actual characteristic of WebSci. One might speculate that we may use more different words in the humanities whereas (some) technology has more repeatable terminology. By supplementing NLP with other techniques such as citation analysis, we can hope to mitigate against any such bias.

The improvement of the methods and the underlying dataset can also lead to more detailed results informing further actions. The NLP technique has already been applied in the Internet Science community, in which the identified application contexts have been used to structure a repository of design methods (Hooper et al., 2014); we hope to complement this with citation analysis in the near future.

Further, future questions include how we might use NLP analysis to understand differences in disciplines according to context: for example, how might a discipline such as sociology appear in a WebSci corpus compared to within a pure sociology corpus? Issues include gaining datasets that are representative of a given domain: for example, the BAWE dataset at Insight has only 111 sociology documents, many of which are short student essays. Of course, insights into this would be strengthened with parallel citation analyses of the different contexts in order to understand the knowledge flows and how they vary across contexts.

Another topic is the impact of shifts in terminology over time, as discussed at CHI (Ibargoyen et al., 2013) or the increasing number of topics a conference deals with over time, as confirmed for Hypertext (Agarwal et al., 2017). We plan to more thoroughly study authors and their effect on the knowledge map and intellectual structure of WebSci. Since Agarwal et al. (2017) have shown that ~50% of authors of web research-related conferences are female, we want to study the relationship between gender and WebSci themes as well as citation networks. Further, it is interesting to investigate whether specific roles related to the conference organisation (e.g., program chair) or frequent publication also affect the knowledge map (Agarwal et al., 2017).

The comparison of knowledge base and knowledge transfer with co-author networks will provide further details about the density and quality of citation networks, also expanding our knowledge about what authors outside the core WebSci community provide input to, or are affected by, WebSci research. The core community of WebSci can be found by calculating the overlap of authors that cite WebSci publications, that publish in WebSci publications, and that, finally, are cited by authors of WebSci publications.

ACKNOWLEDGMENTS

The authors wish to thank the reviewers for their helpful comments. The research leading to these results received funding in part from the Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight).

REFERENCES

Agarwal, S., Mittal, N., & Sureka, A. (2017). A General Overview and Bibliometric Analysis of Seven ACM Hypertext and Web Conferences. International Journal of Web Engineering and Technology. 12(3). doi: 10.1504/IJWET.2017.088376

Åström, F. (2002). Visualizing Library and Information Science concept spaces through keyword and citation based maps and clusters. In

H. Bruce, R. Fidel, P. Ingwersen & P. Vakkari (Eds), Emerging frameworks and methods: Proceedings of the fourth international conference on conceptions of Library and Information Science (CoLIS4) (pp. 185-197). Greenwood Village: Libraries Unlimited.

Barthélémy, M. (2004). Betweenness centrality in large complex networks. The European Physical Journal B - Condensed Matter and Complex Systems, 38(4), 163-168.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 10, P10008.

Bordea, G., & Buitelaar, P. (2010). DERIUNLP: A context based approach to automatic keyphrase extraction. In Proceedings of the 5th International Workshop on Semantic Evaluation (pp. 146-149). Stroudsburg, PA, USA: Association for Computational Linguistics.

Bordea, G., Buitelaar, P., & Polajnar, T. (2013). Domain-independent term extraction through domain modelling. In Proceedings of the 10th International Conference on Terminology and Artificial Intelligence, Paris, France. Retrieved from http://www.insight-centre.org/sites/default/files/publications/tia2013.pdf

Bordea, G. (2013). Domain adaptive extraction of topical hierarchies for Expertise Mining (Doctoral dissertation). Retrieved from ARAN Repository. Available from http://hdl.handle.net/10379/4484

Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. Journal of the American Association for Information Science and Technology, 57(8), 1060-1072.

Chen, C., & Carr, L. (1999). Trailblazing the literature of hypertext: author co-citation analysis (1989–1998). In Proceedings of the 10th ACM Conference on Hypertext and Hypermedia (pp. 51-60).

Cole, J. R., & Cole, S. (1973). Social Stratification in Science. Chicago, IL: University of Chicago Press.

Coulter, N., Monarch, I., & Konda, S. (1998). Software engineering as seen through its research literature: A study in co-word analysis. Journal of the American Society for Information Science, 49 (13), 1206–1223.

Crane, D. (1972). Invisible Colleges: Diffusion of Knowledge in Scientific Communities, Chicago, London: University of Chicago Press..

Culnan, M. J. (1987). Mapping the intellectual structure of MIS, 1980-1985: A co-citation analysis. MIS Quarterly, 11(3), 341–353.

Düzyol, G. , Taşkın, Z., & Tonta, Y. (2010). Mapping the Intellectual Structure of the Open Access Field Through Co-citation Analysis. In IFLA Satellite Pre-conference: Open Access to Science Information Trends, Models and Strategies for Libraries, Crete, Greece. Retrieved from http://eprints.rclis.org/14910/

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., ... & Vespignani, A. (2018). Science of science. Science, 359(6379), eaao0185.

Garfield, E. (2004). The unintended and unanticipated consequences of Robert K. Merton, Social Studies of Science, 34(6), 845-853.

Hall, W., Hendler, J., & Staab, S. (2016). Web Science Manifesto: Retrieved from: http://www.webscience.org/manifesto

Haustein, H., & Larivière, V. (2014). A multidimensional analysis of Aslib proceedings – using everything but the impact factor. Aslib Journal of Information Management, 66(4), 358-380.

Hooper, C. J., Bordea, G., & Buitelaar, P. (2013). Web Science and the Two (Hundred) Cultures: Representation of Disciplines Publishing in Web Science. Web Science, (pp. 162-171). Paris, France.

Hooper, C. J., Hedge, N., Hutchison, D., Papadimitrious, D., Passarella, A., Sourlas, V., et al. (2014). EINS Deliverable 2.3: Whitepaper on recommendations for funding agencies. Network of Excellence in Internet Science FP7-288021.

Hooper, C. J., Millard, D. E., & Azman, N. (2014). Interdisciplinary Coups to Calamites (workshop). Web Science 2014. Bloomington, IN, USA.

Hooper, C., Marie, N., & Kalampokis, E. (2012). Dissecting the Butterfly: Representation of Disciplines Publishing at the Web Science Conference Series. Proc. WebSci 2012 (pp. 137-140). ACM Press.

Hooper, C., Trossen, D., & Surridge, M. (2014). D2.1.1 Repository of methodologies, design tools and use cases. Network of Excellence in Internet Science FP7-288021.

Ibargoyen, A., Szostak, D., & Bojic, M. (2013). The Elephant in the Conference Room:Let's Talk About Experience Terminology. In CHI'13 Extended Abstracts on Human Factors in Computing Systems, Paris, France (pp. 2079-2088).

Jacomy, M. (2011, June 6). ForceAtlas2, the new version of our home-brew Layout. (Gephi) Retrieved from http://gephi.org/2011/forceatlas2-the-new-version-of-our-home-brew-layout/

Jiang B., Endong X., & Jianzhong Q. (2015). A Domain Independent Approach for Extracting Terms from Research Papers. In Proceedings of the 26th Australasian Database Conference, Melbourne, Australia (pp.155-166).

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. American Documentation, 14(1), 10–25.

Kraker, P., Jack, K., Schlögl, C., Trattner, C., & Lindstaedt, S. (2013). Head Start: Improving Academic Literature Search with Overview Visualizations based on Readership Statistics. In Proceedings of the ACM Web Science Conference, Paris, France. Retrieved from http://www.christophtrattner.info/pubs/websci2013_submission_179.pdf

Lopez-Herrera, A., Cobo, M., Herrera-Viedma, E., & Herrera, F. (2010). A bibliometric study about the research based on hybridating the fuzzy logic field and the other computational intelligent techniques: A visual approach. International Journal of Hybrid Intelligent Systems, 7(1), 17-32.

McCain, K. W. (1986). Co-cited author mapping as a valid representation of intellectual structure. Journal of the American Society for Information Science, 37(3), 111–122.

Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. Science, 159(3810), 56-63.

Navigli, R., Velardi, P., & Faralli, S. (2011). A graph-based algorithm for inducing lexical taxonomies from scratch. In T. Walsh (Ed.), Proceedings of the Twenty-Second international joint conference on Artificial Intelligence (pp. 1872-1877). doi: 10.5591/978-1-57735-516-8/IJCAI11-313

Nerur, S. P., Rasheed, A. A., & Natarajan, V. (2008). The intellectual structure of the strategic management field: An author co-citation analysis. Strategic Management Journal, 29(3), 319-336.

Sahal, A., Wyatt, S., Passi, S., & Scharnhorst, A. (2013). Mapping EINS--An exercise in mapping the Network of Excellence in Internet Science. In Proceedings of the 1st International Conference on Internet Science, Brussels, Belgium (pp. 75-78). Retrieved from https://arxiv.org/abs/1304.5753

Small, H. (1973). Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. Journal of the American Society for Information Science, 24(4), 265–269.

Tonta, Y., & Düzyol, G. (2010). Mapping the structure and evolution of electronic publishing as a research field using co-citation analysis. In 14th International Conference on Electronic Publishing, Helsinki, Finland. Retrieved from http://eprints.rclis.org/14695/

Van Eck, N. J., & Waltman, L. (2016). VOSviewer Manual. Retrieved from http://www.vosviewer.com/getting-started#VOSviewer manual

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523–538.

Van Eck, N. J., & Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some wellknown similarity measures. Journal of the American Society for Information Science and Technology, 60(8), 1635–1651.

Wagner, C., Roessner, J., Bobb, K., Klein, J., Boyack, K.,. Keyton, J., Rafols, I. & Börner, K. (2011). Approaches to Understanding and Measuring Interdisciplinary Scientific Research (IDR): A Review of the Literature. Journal of Informetrics, 165, 14-26.

Wang, Z.-Y., Li, G., Li, C.-Y., & Li, A. (2012). Research on the semantic-based co-word analysis. Scientometrics, 90(3), 855-875.

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. Journal of the Association for Information Science and Technology, 32(3), 163-171.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. Science ,316, 1036-1039.

Zhang, Z., Iria, J., Brewster, C., & Ciravegna, F. (2008). A comparative evaluation of term recognition algorithms. In Proceedings of the Sixth International Conference on Language Resources and Evaluation, Marrakech, Morocco (pp. 2108-2113). Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/pdf/538_paper.pdf

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996–2005: Introducing author bibliographic-coupling analysis. Journal of the American Society for Information Science and Technology, 59 (13), 2070–2086.

Zhao, D., & Strotmann, A. (2011). Intellectual structure of stem cell research: a comprehensive author co-citation analysis of a highly collaborative and multidisciplinary field. Scientometrics, 87(1), 115-131.