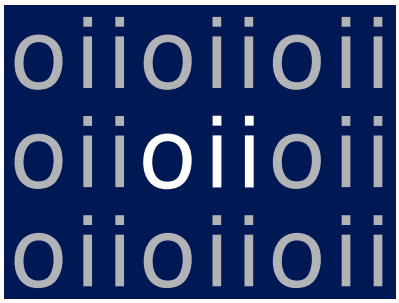


Oxford Internet Institute
University of Oxford

Networks of collaboration and field emergence in 'Internet Studies'

Thesis submitted in partial fulfilment
of the requirement for the degree of
MSc in Social Science of the Internet
at the Oxford Internet Institute
at the University of Oxford.

Momin M. Malik
Balliol College
9,934 words
1 August 2012
[OII Library Edit, October 2012]

Acknowledgements

I would like to thank the faculty of OII, as well as the OII DPhil candidates and my MSc colleagues for an extremely productive year. Thanks to my advisor, Dr Eric Meyer, for timely feedback, and for guidance about informatics literature and methods. Thanks especially to Dr Bernie Hogan for crucial assistance with social network analysis theory, methods, and tools, as well as for instruction around data management, programming, and digital social research. Thanks to Professor Bill Dutton for his previous work in this area, as well as for his encouragement and keen interest in my topic; to Dr Grant Blank for his teaching in statistics and data analysis; to Dr Sandra Gonzalez-Bailon for help in thinking about the importance of having socially grounded and inspired research questions guide computational analysis; to Dr Monica Bulger and Dr Christobal Cobo, with whom discussion about their work with the Network for Excellent in Internet Science provided the inspiration for this thesis topic; and to Professor Ralph Schroeder for discussions that helped further convince me of the power of social constructionist understandings of science and technology. Thanks also to Professor Viktor Mayer-Schönberger for ongoing ideas through the year that provided important motivations for my choices of data gathering and analysis.

Lastly, special thanks to Winson Peng and his team at the Web Mining Lab at the City University of Hong Kong for providing me with the results of their cluster analysis, as well as the results of their Web of Science data collection.

Abstract

Previous work about the emergence and status of the social research area of 'Internet Studies' has focused either on personal perspectives from scholars working in the area, or cross-sectional and semantic analysis of such literature. This work is the first to provide a longitudinal, quantitative, and social examination of the emergence of social science and humanities literature about the Internet from 1990 to 2011, using a co-authorship network constructed from the bibliographic database of the Thomson Reuters Web of Science. I find a critical point in 2000, when a giant component emerges in the co-authorship network. Furthermore, using a novel analytical technique of projecting edges from future years onto nodes of previous years, I find that the growth in coherence of the network subsequent to 2000 was driven not by new authors entering the field, but by the established scholars who were already in the field prior to 2000. The findings have implications for the field self-knowledge of Internet studies as well as the general study of field emergence, and the proposed projection technique is applicable for general longitudinal social network analysis.

Table of Contents

List of Figures

List of Tables

1. Introduction

Recently, scholars of the Internet have reflected on the state of "Internet studies": is it a field? Is it a discipline? What *should* it be: a field, a discipline, or something else?

In the first large-scale, data-driven approach, Peng, Zhang, Zhong, & Zhu (2013, forthcoming) perform a semantic network analysis on data mined from the Thomson Reuters (formerly ISI) Web of Science to find the semantic structure of articles that talk about the Internet. However, they note that they have only the crude measure of the number of articles published about the Internet to identify whether there is a coherent field of Internet studies. Complimenting this approach, I take an approach looking not at the consistency of language, but at the coherence of the social network of co-authorship. Specifically, I conduct a longitudinal analysis to identify the dynamics by which this coherence arose. Drawing on theory from social network analysis, as well as theory from the sociology of scientific knowledge and methodology from bibliometrics,[1] I argue that the proportion of co-authorship links that form the single largest connected component (LCC)[2] in the network has substantive meaning, and that its change over time is informative about the emergence of the field.

The structure of the thesis is as follows:

- In the literature review, I cover first the meta-literature about Internet studies, then the various ways in which previous studies have interpreted co-authorship networks, and lastly the technical literature pertinent to the emergence of the 'giant component', which is when the LCC accounts for a plurality of the network.
- In methodology, I explain the decisions I made in collecting, processing, and analysing bibliographic records from the Web of Science (WOS).
- In findings, I discuss evidence of the coherence of the field of Internet studies by showing the change in the proportion of edges in the LCC over time, and look at how the network coalesces.

[1] A field devoted to quantitatively analysing publication metadata.
[2] A *component*, in network terminology, is an interconnected cluster. Components are like 'islands' in the network; it is impossible to get from one to another through the edges in the network. A single, unconnected node (i.e., a 'trivial' component) is called a *singleton*.

- In the discussion, I interpret the evolution of the dynamics: I argue that after 2000, the key growth in the field was driven by *consolidation*, defined as new edges between incumbent nodes, rather than *spread*, defined as edges created through the entrance of new nodes.
- In conclusions, I discuss the implications of my findings both for the self-knowledge of the field of Internet studies and for the general analysis of networks, as well as directions for future research—both substantive and theoretical—opened by the findings.

With this study, I endeavour to make both a theoretical contribution and a substantive one. The theoretical contribution I make is in proposing a modelling technique where the *edge* structure of a given time slice is projected onto the existing *nodes* of an earlier time slice, as a way of exploring the dynamics of social network growth. The substantive contribution I make is in adding empirical findings to ongoing reflection within the field of Internet studies.

2. Literature Review

2.1. Internet studies

The *International Handbook of Internet Studies* (Springer) was published in 2010, followed shortly by *The Handbook of Internet Studies* (Wiley) in 2011. These are being joined by the forthcoming *Oxford Handbook of Internet Studies* (OUP) in 2013. For so many different publishers to try and capitalise on a field at the same time is a sure sign of contemporary relevance, as well as a sign that there are enough researchers with enough output to fill up multiple unique volumes. Other signs of the field's established nature include the number of degree programs and research centres worldwide: Mazar (2010a; 2010b) counts 33 degree programs and 46 research centres and institutes engaged in related research. Internet-specific journals such as *New Media and Society* and *Cyberpsychology, Behavior, and Social Networking* are thriving, and the proportion of Internet-related articles in well-established journals such as the *Journal for the American Society of Information Science & Technology* has been steadily rising (fig. 1).

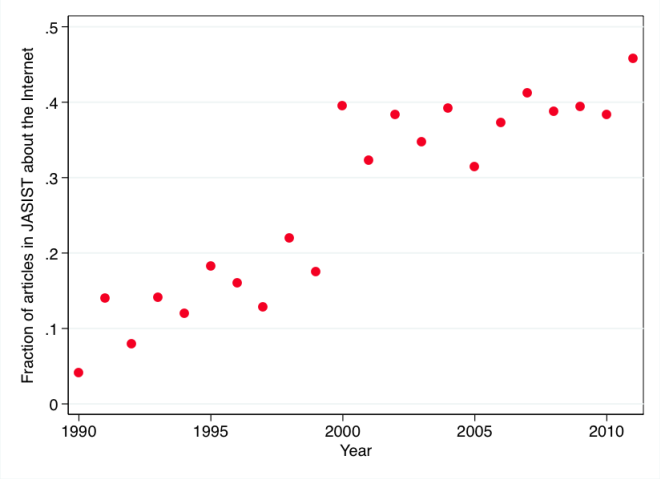


Figure 1. Proportion of articles in the *Journal of the American Society for Information Science & Technology* (prior to 2001, the *Journal of the American Society for Information Science*) about the Internet. The total number of JASIS(T) articles is from a WOS query. For the identification of the number of Internet articles, see methodology below. Note the conspicuous jump at 2000.

There is little scepticism that Internet studies exists, but there is debate about the nuances of its status and the timeline of its emergence. There was wide consensus in 2005 against the notion of Internet studies being a discipline (Baym, 2005; Jones, 2005; Markham, 2005; Sterne, 2005), and in favour of seeing it as a field, a position that remains today (Dutton, 2013, forthcoming).[3] There is also debate about when exactly it began: Dutton (Ibid.) critiques the view that Internet studies began with the first conference of the Association of Internet Researchers in Lawrence, Kansas in 2000, instead taking a 'long view' that sees Internet studies as continuous, with earlier strands of research into computer-mediated communication. Silver (2004, p. 57) identifies a more recent beginning, placing the academic beginnings of Internet studies around 1995-1997, when scholars avoided the disciplinary foci and slow turnaround time of journals by publishing several anthologies relating to 'cyberculture'. Silver argues that these anthologies, representing a variety of theoretical perspectives, were the beginning of the field. Wellman (2004, p. 125) also calls this period as the "first age of Internet studies"; however, he sees it as devoted largely to naïve cyber-utopianism/cyber-dystopianism, and identifies genuine research beginning in the "second age of Internet studies" around 1998. This 'age' too was not yet theoretical but descriptive: it was "the low-hanging fruit [gathered by] analysts using

[3] Baron (2005) argues that while fields and disciplines are ostensibly distinguished by fields being about the nature of the problems being investigated, and disciplines being an academic identity made up of common tools and a common community, the distinguishing feature between the two is less about "the nature of the academic enterprise" or academic identity and more about the "institutional aspirations" of practitioners: generally, only disciplines command budgets (including receiving office space, secretaries, and funding for graduate students), whereas fields may be a way for academics to achieve "an intellectual escape from their academic home, which sometimes dismissed their work as falling outside the disciplinary pale." Baym (2005, p. 229) offers the following definition: "Disciplines have clear organizational forms. There are departments, research centers, office spaces, support staff, letterhead stationary, and perhaps even endowed chairs. Internet research has none of these."

standard social scientific methods—and some concepts—to document the nature of the internet" (Ibid., p. 127).

Regardless of whether the anthologists or the documenters should get recognition for being the first in the field, it was around the second period that there emerged a great deal of diverse activity. There arose Internet-specific journals such as *Cyberpsychology & Behavior* (since 2010, *Cyberpsychology, Behavior, and Social Networking*), started in 1998, *New Media and Society*, started in 1999, and (not as major but an example of extreme specialisation around the Internet) the *Journal of Medical Internet Research* (JIMR), also started in 1999. There was the founding of research centres including, out of those still active, the Berkman Center for Internet & Society at Harvard Law School in 1998, the Center for Digital Discourse and Culture at Virginia Tech 1998, the Centre for Internet Research at the University of Aarhus (Denmark) in 2000, and the Oxford Internet Institute in 2001 (list from Mazar, 2010b). Although Internet research occurs within many professional organisations, a devoted professional association, The Association of Internet Researchers (AoIR), began in 1998 (McLemee, 2001), with its first meeting in 2000 in Lawrence, Kansas. Brandeis University began offering the first degree program in Internet studies in 2001 (McLemee, 2001). Thus, within the span of only three or so years, a number of researchers—by founding research centres, journals, and a professional organisation—were self-consciously carving out a research area for the social study of the Internet.

A second, minor inflection point seems to be about half a decade later: the NEXA Research Center for Internet and Society at Politecnico di Torino (Italy) was founded in 2006, and shortly before in 2004, two reflections on Internet studies (Wellman, 2004; Silver, 2004) appeared in the fifth anniversary issue of *New Media and Society* (2004). Similarly, a special issue of *The Information Society* in 2005 (Vol. 21, No. 4) was devoted to articles questioning the existence and nature of Internet studies, including treatments grounded in the sociology of scientific knowledge (Hine, 2005), social construction (Markham, 2005), and science and technology studies (Monberg, 2005). I avoid reviewing these as my current quantitative approach is theoretically motivated and justified but is a separate project from a purely theoretical one. Suffice to say, most of these authors emphasise the priority of community: the status of Internet studies has less to do with its intellectual content than the aspirations and self-identification of its practitioners—disciplinary or field cohesion is at its core a social, rather than intellectual, construct (even if its expression is intellectual).

The most recent treatments of Internet studies look not at whether it is a field or not, but at its subdivisions. While Rice (2005) conducted a preliminary semantic network analysis, it was only on session titles and paper titles and abstracts from the 2003 and 2004 meetings of AoIR, and thus has little longitudinal power and is a very limited sampling. Dutton (2013, forthcoming), from his experience in the field, proposes that Internet studies has three main objects of study (technology, use, and policy), each cross-referenced with three key issues (who shapes, why, and with what implications). In a different approach, Peng et al. use semantic network analysis on over 25 thousand Web of Science records to identify four main topic clusters (and eleven total subclusters): e-health (generic applications, specific behaviors); e-business (acceptance studies, management & Internet, marketing & Internet); e-society (social interactions & Internet, law/policy & Internet, communication & Internet); and human-technology interaction (psychological processing & Internet, web search/e-library, e-learning).

However, the basis of Peng et al.'s basic claim that the field of Internet studies exists and is coherent is based on sheer numbers (asking, "Is the volume of research output sufficient for Internet studies to be considered as a field within the social sciences?"), and not structure. In comparison with other topic searches, they find that "'Internet' ranks third among the seven fields, below 'environment' (38,719) and 'society' (27,357) but above 'culture' (26,937), 'economy' (20,596), 'politics' (20,165), and 'globalization' (7,457)", yet there is no guarantee that these topic searches represent coherent fields in themselves. Semantic network analysis, and basic frequency analysis, is not able to make any comment about the existence of a field; that requires another method.

By reviewing points of relevance only in the meta-literature about Internet studies, I do not mean to imply that the field evolved and matured by its own momentum. The growing social and commercial visibility of the Internet is likely the main driving factor for the field's growth: after all, much of academic work (and training) is constrained by available funding, and available funding—understandably—is constrained by what is currently understood to have societal relevance. I do not aim to carry out an explanation of the growth of Internet studies, only the manner in which it grew.

2.2. Co-authorship[4]

Collaboration can take many forms, but perhaps the most concrete is that of co-authorship. When multiple authors publish together and generate bibliographic metadata, it is an

[4] The following text is an adaptation of the literature reviews of the two option papers I submitted in Hilary Term.

unambiguous evidence of a social tie: even if one author contributed nothing to a given paper, perhaps giving it a rubber-stamp approval without even reading it, such an arrangement is still evidence of a social tie. But the social interpretation of co-authorship networks has thus far been limited.

Physicists and computer scientists typically use co-authorship networks as 'model systems' on which to evaluate data mining algorithms (Gehrke, Ginsparg, & Kleinberg, 2003), or community detection algorithms (e.g., Newman 2000, 2001a; Jin, Girvan, & Newman, 2001; Barabási et al., 2002). Co-authorship networks are very attractive for these purposes because they are a relatively clean data source and because existing bibliographic databases yield copious amounts of data. However, such uses do not have substantive meaning: they only try to find ways to replicate human understandings (such as a way to detect clusters of subspecialties within the co-authorship networks). Developing metrics and data mining techniques is certainly an important task, but it is different from finding ways to analyse network that give social insight.

There is also bibliometric and scientometric literature that looks at ways to use co-authorship for assessment. Franceschetа and Costantini (2010) find a connection between quality (as judged by an assembled panel) and co-authorship between authors of heterogeneous affiliation. Abbasi, Altmann, and Hossain (2011) look at how centrality measures in a co-authorship network correlate with academic 'performance' (as measured bibliometrically). Other literature looks at patterns in co-authorship to make comments about the working of academia: for example, literature on international co-authorship notes a rise in the prevalence of co-authorship across national borders (Leydesdorff & Wagner, 2008; Persson et al., 2004), but that co-authorship within national boundaries is still 10-50 times more likely than international co-authorship (Hennemann, Rybski, Liefner, 2012).

Such bibliometric studies, however, tend to be tied only to speculation about the underlying dynamics. For example, Leydesdorff and Wagner (2008) make the argument, "Patterns in international collaboration in science can be considered as network effects, since there is no political institution mediating relationships at that level except for the initiatives of the European Commission." What this fails to recognise is how trans-national communities of academics can be powerful bodies for mediating relationships, and how academic institutions often play international roles (not least in the circulation of academics between universities). I would argue that such speculation confuses network *effects* with network *signatures*. I

contextualise patterns found through network analysis as only signatures of underlying dynamics, rather than the dynamics themselves.

A study falling between assessment and substantive social study is Lambiotte and Panzarasa (2009), who ask what network properties represent for the production of scientific knowledge. They note that dense, modular structures represent specialisation, and that this has implications for scientific work: specifically, they examine the "trade-off between social cohesion and brokerage by investigating the conditions under which scientists can enhance their performance by collaborating with others within or outside their own communities". While moving towards greater substance in its theoretical interpretation, I would argue this work also confuses the network signatures for decision-making by nodes in the network. Academics do not make co-authorship decisions based on a desire maximising performance—funding opportunities, departmental politics, institutional location, and other exogenous factors likely play a greater role.

One work that gives substantive interpretation to co-authorship networks is Bettencourt, Kaiser and Kaur (2009). Their work relies on a common feature of networks: the giant component, which I will first review in its own right.

2.3. The giant component

An interesting feature of large networks, as observed by a number of physicists and computer scientists (Newman 2000, 2001a; Barabási et al., 2002; Leskovec, 2008; Leskovec, Lang, Dasgupta, & Mahoney, 2008), that once networks reach a certain size they tend to undergo a 'percolation transition[5] that leads to the emergence of a 'giant component': this is a LCC that accounts for a majority of the edges in the network (and, usually, a majority of nodes in the network as well).[6]

For academic co-authorship networks, the rise of a giant component is rather unintuitive. We imagine academic work to be structured in silos, with scholars collaborating only with a small and self-contained set of colleagues. Yet large enough co-authorship networks indeed have a

[5] A mathematical term for crossing a critical threshold for the probability of edge formation over an underlying lattice, whereupon the network gains large clusters and long-range connectivity.
[6] The use of edges, rather than nodes, as the criterion is because edges tell about density and structure, not just size. Note that considering the fraction of edges in the LCC excludes singletons (which don't contribute to network dynamics anyway), and also that it discriminates against smaller connected components (as the maximum number of possible edges between n nodes is bounded by the n^{th} triangular number, and social networks tend to be sparse anyway; Newman, 2010).

giant component, implying that there is greater underlying coherence than might appear to be the case from qualitative impressions.

Still, there is ambiguity about whether the giant component is anything with substantive meaning. Physicists and computer scientists regard it as a mathematical property related to size and the network density. After all, giant components arise in simulated random networks, which is not hard to understand: as the network size or density grows, the possible paths by which any two nodes be connected grows exponentially, and with it the probability of there coming to be a path connecting the two nodes. Even more powerfully, if there are two disconnected large components, a link from any one node in one component to any one node from the other component would merge the two components. To take a real-life example, every author in the LCC I will identify for the co-authorship network of Internet studies has a well-defined Erdős number,[7] as the sociologist Barry Wellman published with statistician Ove Frank (Wellman, Frank, Espinoza, Lundquist, & Wilson, 1991), who had an Erdős number of 2. All it took was one link to connect a giant connected network of sociologists to a giant connected network of mathematicians. Nor is Wellman the only link; the mathematician Joseph O'Rouke (who also had an Erdős number of 2) publishing with sociologist Eszter Hargittai (Feigenbaum, Hargittai, & O'Rourke, 1994) is another,[8] and there may well be more.

Still, the example of sociologists having Erdős numbers points to an important difference between simulated random networks and actual networks: random networks undergo percolation transitions, whereas actual (co-authorship) networks have connections created between people when they work together. Wellman publishing with Frank, or Hargittai publishing with O'Rouke, is not a random output from a probability function, but a human connection whose existence came into being from a deliberate process. As networks expand, the way in which networks can be interconnected grows exponentially; but to make a connection is still a specific, and meaningful social event.

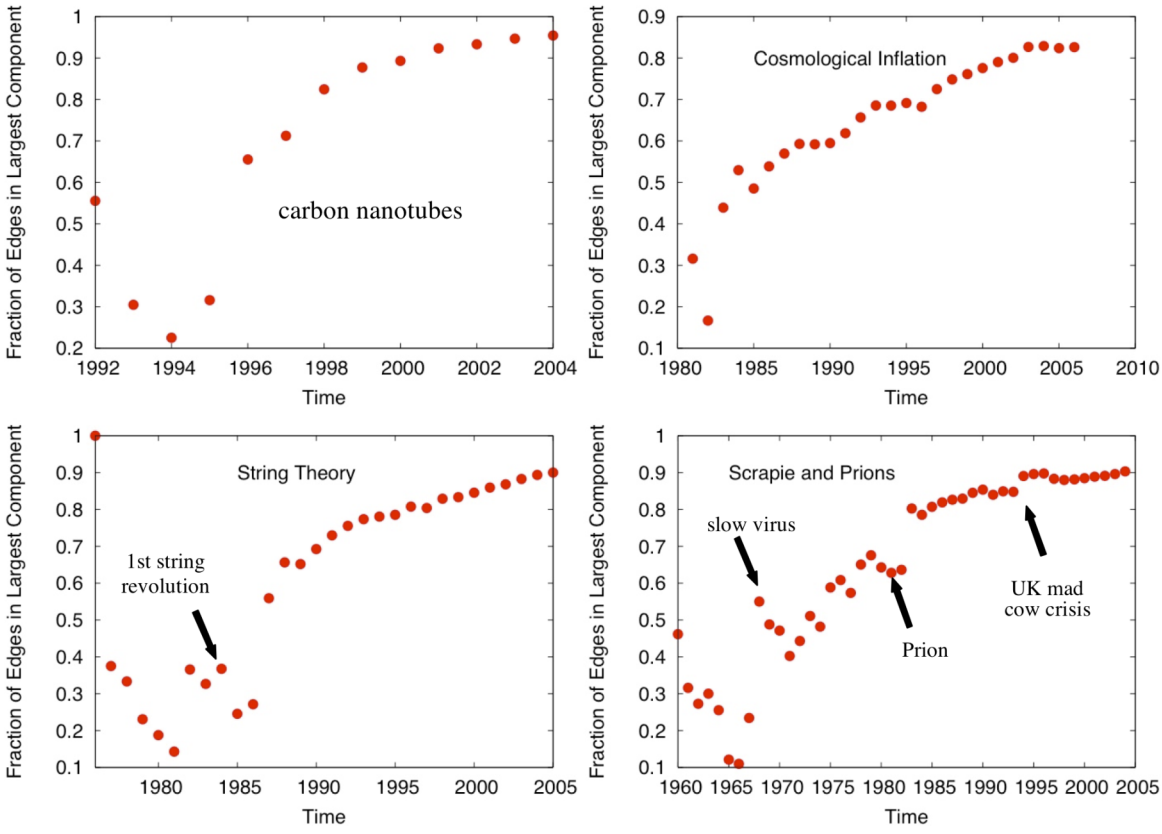
But even if we stop viewing the emergence of a giant component as a mathematical artefact, the giant component itself may not have much sociological meaning. Just because a co-authorship network of sociologists and mathematicians forms a connected component does not mean that

[7] Paul Erdős (1913-1996) was a prolific and eccentric mathematician, and mathematicians created the concept of an Erdős number as a tribute to him. The (minimum) degrees of co-authorship separation between a person and Erdős is a person's Erdős number (Erdős himself defined as having an Erdős number of 0). Having a low Erdős number is often a point of academic pride.
[8] http://www.eszter.com/erdos.html

they form a coherent community: but separating existence from coherence is a tricky mathematical prospect.[9] The real difficulty of interpreting connected components relates to the need for a 'null model': a background against which to interpret the significance of results.

The study by Bettencourt, Kaiser and Kaur (2009) is key in that it takes the LCC and finds the behaviour of its growth varies across fields. Specifically, it finds one pattern for 'successful' fields, and a different pattern in a selected 'pathological' field, cold fusion, that failed to become established despite hundreds of publications—cold fusion in effect becomes a null model.

Their study analyses the co-authorship networks of eight scientific fields from their inception: superstring theory, cosmic strings and other topological defects, cosmological inflation, carbon nanotubes, quantum computing and computation, prions and scrapie, H5N1 influenza, and cold fusion. They find a number of measures by which cold fusion differs from the other five fields. One is the number of new authors; the five successfully emergent fields show a steady rise, whereas cold fusion shows a peak before dropping down to less than 50 per year. Another is that cold fusion alone did not show a steady rise in the fraction of edges in its LCC (fig. 2).



[9] This is the idea behind 'community detection' algorithms, but community detection is still a developing research area with a great deal of technical uncertainty. See discussion in Appendix B.

Figure 2(a). Reproduction of Fig. 6 from Bettencourt, Kaiser and Kaur (2009, p. 218).[10] "Time series for the fraction of edges in the largest component. The increase in the fraction of edges in the largest component of emerging fields suggests that the introduction of new concepts and techniques leads to a topological transition where most scientists in the field become connected by ties of collaboration."

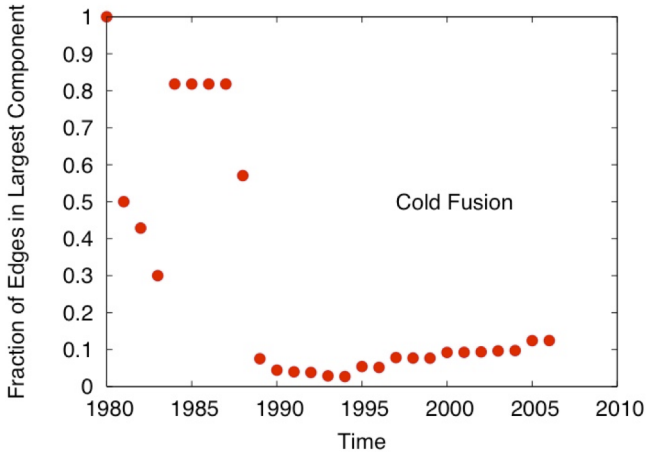


Figure 2(b). Reproduction of Fig. 7 from Bettencourt, Kaiser and Kaur (2009, p. 218), "Field development and topological critical behavior… All successful fields [above, 2(a)] display the same approximate critical behavior… Fields without an established (and shared) set of concepts and techniques, such as cold fusion [below, 2(b)], do not display a topological transition."

The key theoretical link is made with Kuhn's (1970) theory of scientific revolutions, and specifically his hypothesis about a practice of "normal science". Bettencourt et al. argue that for a scientific practice to become enshrined, it requires a stable community of practitioners—thus justifying the use of the social connections of co-authorship networks, rather than informational connections of citation networks or 'artificially made similarity-based connections' such as word co-occurrence networks (Yan & Ding, 2012, p. 1314). The failure of cold fusion to develop a coherent community, as would be represented by a large connected component in the co-authorship network, is tied into its failure to become an established field. That is, the network signature ties into underlying dynamics, and has substantive sociological meaning. Cold fusion shows that the rise of a giant component is not inevitable, and that it may be taken as a signature of a maturing field.

Bettencourt et al. provide a null model of a field that did not coalesce, along with a theoretical explanation, and thus forms the key background for my study. However, there are two key differences between their study and this thesis. First, Bettencourt et al. studied sciences, using a cross-database search, rather than studying social sciences and humanities through only the

[10] Figures from Bettencourt et al. taken from pre-print at http://web.mit.edu/dikaiser/www/BKK.Topological.pdf. Used by permission of David Kaiser. Page references are to the version published in the *Journal of Informetrics*, *3*(3).

WOS Social Science Citation Index and Arts & Humanities Citation Index as I do. This is important both because co-authorship is still less frequent in the social sciences, although it is increasing (Wuchty, Jones, & Uzzi, 2007), and because the dynamic that Bettencourt et al. identify as behind the growth of a giant component, postdoc positions (deliberated invented to generate circulation of young scientists), is not institutionalised outside of the sciences. Hine (2005, p. 240) points out a strand of literature arguing that Kuhn's ideas of paradigm shifts do not apply to the social sciences. However, she also offers, "Probably the key insight to draw from Kuhn… is that knowledge and the ways to produce it are validated within the community". We do not require Kuhn's overall model of scientific development to build on the insight that intellectual practice is a fundamentally *social* phenomenon, and that looking at social structure can be more powerful for understanding intellectual dynamics than is looking at intellectual structure.

Second, Bettencourt et al. take well-defined fields, and they take the entirety of their literature (both journal articles and conference proceedings, aggregated across several bibliographic databases). The sheer spread of where material about the Internet is published is a challenge not faced in other fields (e.g., Heersmink et al., 2011, p. 242). I am taking the somewhat tautological approach of using a topic search to identify an emergent field, and then using the results of that search to claim that the field exists. However, as I will argue, while a single time slice cannot tell us about coherence, a time series revealing the dynamics of emergence can. The network is not coherent when it first emerges, but slowly gains structure as time goes on; this process of change, from a network that lacks coherence to one that has coherence, is meaningful. I also argue with regards to the scope of sampling that the network I identify is certainly not the entirety of the co-authorship network of Internet studies, but it is certainly a large proportion of it, making the findings have the significance of brute size.

3. Methodology

3.1. Data source

My data source is the Thomson Reuters (formerly ISI) Web of Science (WOS), which consists of a number of linked databases across the sciences, social sciences, and humanities, and across journal articles, conference proceedings, and books. It calls itself "indisputably the largest citation index available",[11] with 10.7 million records in social science and arts and humanities

[11] http://wokinfo.com/realfacts/qualityandquantity/

over 6,877 journals, [12] including complete citation coverage from 1900.[13] The use of WOS and choice of it over alternatives is well supported in literature. Furthermore, the study whose data collection approach mine most closely resembles, Peng et al., uses the WOS.

Unfortunately, the WOS (currently in version 5.5[14]) does not document its version history,[15] and specifically, if and when there have been any changes in methods of data storage, processing, or management (although it does document its journal selection process[16]). One of my significant findings relates to a jump in the data at 2000, and I would like to have confirmed that there were no changes in the WOS system that might account for the shift. However, considering that the WOS originated as an academic project,[17] is a major tool for the academic community, and metrics calculated from it are even used to inform hiring decisions in academia,[18] any interruption in the data consistency at 2000 would have far-reaching repercussions. Indeed, it would be a grave oversight of Reuters to not make sure the database is consistent across its indexed years. Thus, I acknowledge the possibility that my findings are an artefact of how the WOS operates, and have no way of disproving this, other than to say that institutional factors make it extremely unlikely.

There are important limits to what I can conclude from information gathered from the WOS. While the WOS claims to completely index the sources it does cover,[19] it does not cover every journal.[20] There are also some questions about the process of the growth of the WOS, as every year about 2,000 journals are considered for inclusion, of which 10-12% are included, but this might be slower than the overall growth of scholarly output and this might be a problem (Larsen & von Ins, 2010).

But more profoundly, while social community may be the defining feature of a field of study, a community will not be completely captured in co-authorship networks. There are a huge number of informal ties created at conferences and workshops, by visiting programs, and within departments and universities, and even formal collaborations in research projects (Ribes & Bowker, 2008) that nonetheless do not lead to co-authorship. Yet, considering that the LCC I

[12] Adding the social science and arts and humanities numbers from http://wokinfo.com/realfacts/comprehensive/
[13] http://wokinfo.com/realfacts/covertocover/
[14] http://images.webofknowledge.com/WOKRS55B6/help/WOS/hp_whatsnew_wos.html. Leydesdorff, Carley, and Rafols (in press) discuss some salient new features of WOS v5.
[15] The closest there is to version documentation is http://wokinfo.com/about/whoweare/ or http://images.webofknowledge.com/WOKRS55B6/help/WOS/hp_whatsnew_wos.html, but they lack detail.
[16] http://thomsonreuters.com/products_services/science/free/essays/journal_selection_process/
[17] http://wokinfo.com/about/whoweare/
[18] http://wokinfo.com/realfacts/experienced/
[19] http://wokinfo.com/realfacts/covertocover/
[20] http://thomsonreuters.com/products_services/science/free/essays/journal_selection_process/

identify is in excess of 70,000 individual authors,[21] I again argue that its sheer size represents a meaningful (even if non-random) section of the community of Internet studies.

3.2. Data collection

I began by following the basic pattern of Peng et al.'s data collection. They used six query words (Internet, web, cyberspace, cyber-space, online, and on-line) in the Web of Science "topic" field (which searches through titles, keywords, and abstracts), searching the Social Science Citation Index (SSCI) and Arts & Humanities Citation Index (A&HCI) between 2000 and 2009. They limited article langue to English, and document type of scholarly journal articles, and retrieved the information of 27,340 relevant articles. From this result, they conducted additional data cleaning (for example, eliminating hits for "web of science" that came from the "web" search term) and ended up with 25,685 records (Winson Peng, private communication, 26 June 2012).

I departed in several key ways that are appropriate to my methodology. The most important difference is that Peng et al. were conducting a semantic network analysis, meaning that irrelevant articles not really about the Internet caught by the search query would contaminate the subsequent analysis. However, my project makes a co-authorship network, and I specifically look at the LCC. This means that if my search query nets irrelevant articles, they likely won't be part of the LCC, and hence won't matter. Thus I can safely 'overshoot', and in fact should aim to overshoot as much as possible, as false positives will have negligible impact whereas false negatives might fracture the network I create.

I used the following WOS query:

> TS=(internet OR cyber* OR online OR "on line" OR web* OR google OR facebook OR twitter OR myspace OR youtube OR ebay OR wiki* OR *blog* OR "digital divi*" OR "e book*" OR ebook* OR "e business*" OR ebusiness* OR "e govern*" OR egovern* OR "e learn*" OR elearn* OR "e market*" OR emarket* OR "e mail*" OR "electronic mail*" OR email*)
> Databases=SSCI, A&HCI, CPCI-SSH, BKCI-SSH Timespan=1990-2011
> Lemmatization=On

This query yielded 114,079 hits at the time it was made.[22] For details about how I arrived at this query, and the reasons for departing from the query of Peng et al., see Appendix A.

[21] Note that this figure comes from both the Social Science Citation Index, as well as the Arts & Humanities Citation Index (following the methodology of Peng et al.) of the WOS, and thus includes more than social scientists. Also, if there are any computer scientists, statisticians, or physicists who co-author papers with social scientists (or co-author their own social science papers), they would be included in this component as well (for example, notable physicists Albert-László Barabási and Mark Newman are both in the giant component).

A note on ethics: Data-mining from the WOS is a well-established tool for academic research and assessment. Furthermore, all data collection and handling complies with the WOS terms of use:[23] as the amount of data I downloaded "would not have significant commercial value of its own" and "would not act as a substitute for access to a Thomson Reuters product for someone who does not have access to the product", it qualifies as an "insubstantial portion" of the WOS and hence a "reasonable amount" of downloaded data. Nor have I shared the data.[24] Beyond the terms of use, since the data is not private information, but the data trail created by work put by scholars into the academic domain, there are no ethical concerns with the use of the data.

3.3. Data processing

While the WOS does have an API, it only allows access to a "lite" version with limited fields. For the full number of fields, WOS only allows downloads of 500 records at a time. I manually downloaded some 250 such files, and imported them into an SQL database for management. The latest version of the WOS (v. 5) includes 54 fields (see Table 3 in Appendix B). The main two fields I considered were AU (Author) and PY (Publication Year), although for background consideration I used a number of other fields (for example, I used the journal name to make fig. 1 above, the title field to check correspondence with Peng et al.'s data set, etc.).

I did the bulk of my data processing and management in Python 2.6. I have uploaded all my major scripts to a Github repository[25] (under only my candidate number), for readers to examine if they so wish.

Using Python, I took exported columns from the SQL database and wrote them to the GraphML (http://graphml.graphdrawing.org/) format. During the data processing, I eliminated singletons (i.e., ignoring any row whose AU field lacked a semicolon, the sign of multiple authors) to reduce filesize, although I did count the absolute number of authors (singletons included) with another script. I expanded the set of authors of a given article through a combinatorial script to turn co-authorship instances into a network edgelist, while generating a

[22] The WOS database seems to be constantly growing, as subsequent repetitions of this query yield a few more hits. Because WOS does not have a field to indicate when a record was added, it is impossible to systematically track this.
[23] http://wos.isitrial.com/policy/Policy.htm
[24] Neither did Peng et al. share data with me in an inappropriate way. The data they sent was only a two-column spreadsheet consisting of article title in one column and Peng et al.'s own semantic classification code in the other. I reconstructed the rest of the bibliographic information from matching the title field from the results of my own query.
[25] https://github.com/559755/Scripts-for-MSc-Thesis--Networks-of-collaboration-and-field-emergence-in--Internet-Studies--

sorted and unique list of authors. In the GraphML file, I gave every edge an attribute for the year in which it was formed, as well as giving every node a node attribute for the earliest year in which it appears in the network (taking all the years in which the author was a co-author on a paper, and taking the minimum).

While co-authorship networks are cleaner as a data source than almost any other type of social network, there is still the problem of two people being falsely conflated as the same person, or one person falsely been identified as two people, as author names are not always consistently across all publications. One common problem that is easy to fix is that surnames like "van den Besselaar" or "McPherson" may be inconsistently capitalised across publications, and some data processing methods (such as Python dictionaries) are case sensitive; to address this, I also converted all author names to a uniform letter case. But there are other inconsistencies that are practically impossible to systematically address. For example, Newman might be Mark Newman, Mark E. J. Newman, M. E. J. Newman, MEJ Newman, M Newman, etc. To quantify the extent of this problem, Newman (2000, 2001a) carried out two analyses: first, using all initials of each author alone with the surname (which will sometimes misidentify the same person as two people), and second, using only the first initial of each author with their surname (which will sometimes conflate two authors) to get upper and lower bounds on the expected precision. In his data set, the Los Alamos e-Print Archive, he found the difference between these bounds was 14%. However, the e-Print Archive is a messier data source than the at least partially curated databases of the WOS, so it is reasonable to assume that the flaws I have are less than this bound. While the WOS does have a field for full author name, AF, I used the AU field (which is last name, comma, first and middle initial) rather than aim for an upper bound to the number of authors.

I then imported the GraphML file to NetworkX (http://networkx.lanl.gov/) for analysis. The resulting network had a total of 125,192 nodes and 369,328 edges. I first eliminated multiple lines, keeping the line corresponding to lowest year, as well as eliminating self-loops. This cut down from 369,328 edges to 322,586 (with 46,568 cases of repeat co-authorship, and 87 cases of 'self-co-authorship'[26]). Then, I dropped all but the largest connected component. This left 73,736 nodes (58.90% of the network, excluding singletons, or, including singletons, 50.20% out of the 146,885 total nodes) and 254,943 edges. Every node has, as a node attribute, the year it

[26] Self-loops are impossible; hence, they indicate flaws in data. To see whether the flaw was in my processing or in the raw data, I checked some cases of such supposed self-authorship. I found papers with a few dozen authors having the same author listed twice (or, alternatively and less likely, there are two co-authors with the same last name and same first initial). These are thus flaws in the raw data that I collected from the WOS, but it is a negligible amount: 87 self-loops in the total network, 62 in the LCC.

entered the network, and every co-authorship edge has a 'weight' that corresponds to the year in which it was formed.

3.4. Data analysis

For my findings, I focus on the longitudinal evolution of the largest connected component (LCC). First, I replicate the measures of Bettencourt et al., looking at the number of new authors per year, and the rise over time of the fraction of edges in the LCC. Second, I introduce novel measures to get at underlying network dynamics.

Looking at the number of new authors per year raises a problem. Bettencourt et al. look at well-defined fields, whereas I use the results of a topic search and claim that they represent a field. Should I look only at the new authors per year in the LCC? Or, the new authors per year in the entire network? As I argue above, the LCC without a doubt represents an intersection with a large body of what could be called 'Internet studies'. Looking only at the LCC has the advantage of eliminating any authors not part of the community of Internet researchers, as any author who is not studying the Internet will almost certainly not be part of the LCC (there are likely some authors who publish pieces of related literature who wind up in the LCC, but they would be the vast minority). However, the converse is not true: there will be Internet studies researchers who fall outside the LCC (as an example, Sandra González-Bailón—whose 2011 Twitter paper is definitely an example of Internet studies—is in a component of 10 authors and not part of the LCC). Then, is it worth avoiding false positives (non-Internet studies outside the LCC) at the cost of having false negatives (the Internet studying authors outside the LCC)? I argue that focusing on the LCC is the more relevant measure: since neither every author in my WOS query nor every author in the LCC covers every Internet researcher, it is better to look only at the dynamics of the internally consistent LCC. For new authors, I will do all three analyses (all nodes, singletons excluded, LCC only) because it is relatively easy (and they all turn out to show the same pattern).

But for looking at the rise of the fraction of edges in the LCC over time, I look only at the LCC; i.e., use the LCC at 2011 to bound the network, such that the fraction of edges in the LCC at 2011 will by definition be 1. Then within that I look at how the 2011 LCC grew over time. This is a sort of 'normalisation'; as Bettencourt et al. found (p. 218), there is no absolute number of the fraction of edges in the LCC that signals an established field, that the 'critical

value' is different for different fields, considering the edges outside the LCC would result in a lower absolute fraction per year but the same relative proportion and change over time.[27]

In the second part of my analysis, I will look at network dynamics. There are only two dynamics by which the proportion of edges in the LCC can increase over time: first, with the addition of new nodes that link together previously disconnected components, and second, with edges forming between existing nodes to link components. There will always be a combination of these two dynamics, but their relative contribution to increased coherence changes over time. I call the first dynamic "spread", since it relates to the network becoming more coherent through new entrants, and the second dynamic "consolidation", as it is creating greater density in the network of already existing nodes. The sociological meaning of each of these dynamics is clear: is it new entrants that are making a field coalesce, or is it increasing ties between incumbents that is doing so?

As a first step, I formalise the measure used by Bettencourt et al. (2009). I define "coherence" as *the fraction of edges in the largest connected component*. As an equation:

$$Coherence(t) = \frac{Edges_t(LCC)}{Edges_t(Total)}$$

Note that for my analysis, I take 'Total' to be the subset of the nodes of the 2011 LCC present at time *t*, and 'LCC' to be the largest connected component within that network at time *t*. Also note that coherence is only meaningful plotted over time.

Next, in order to study the extent of consolidation, I propose a novel technique: to project the *edges* from a later time onto the *nodes* of an earlier time, and measure the change in coherence. If the change is positive, I call the dynamic "consolidation", and if negative, "spread".[28] Note that projection is equivalent to taking the network at a given time, removing all nodes that entered on or after a given year, and measuring the coherence (this is how I scripted the computation).

[27] This is not entirely accurate; in the full network, the LCC at 1990 is a component of 9 nodes that does not end up in the final giant component. Thus it is possible I would get slightly different numbers. But since the other connected components don't 'matter' in the end, it is appropriate to exclude them, and I did not investigate further the possibility that the LCC in certain years would not wind up in the 2011 LCC.
[28] I didn't find a good way to formalise consolidation, so I give only the verbal description of what it is.

Edge projection is a way to measure the amount to which new edges form between *incumbent* nodes. I will hypothesise that up until a certain point in time, the network was growing through spread, i.e. a projection of nodes at time $t_{m>n}$ onto time t_n will not result in a marked rise in coherence, even for t_{max}. However, after a certain point, while spread will still continue, the critical mass will be achieved such that the connections between incumbent nodes will be what drove the growth of coherence.

There are a number of analyses that I considered but ultimately rejected, including:

- Use of databases other than the WOS
- Using the Book Authors (BA) and Book Editors (BE) fields along with AU
- Centrality measures
- Network statistics with actor-based modelling
- Regression to predict co-authorship based on discipline
- Comparing communities to the clusters of Peng et al.
- Using ready-made tools for WOS data analysis
- Network visualisation
- Citation network analysis
- Network decay

I rejected these for a variety of reasons, including that they lacked theoretical relevance, were computationally intractable, had too many problems with data consistency, or were insufficiently developed in the mathematical literature. See Appendix B for detailed explanations.

4. Findings

4.1. Network growth

The first observation to make is the growth in size of the network (i.e., the number of authors).

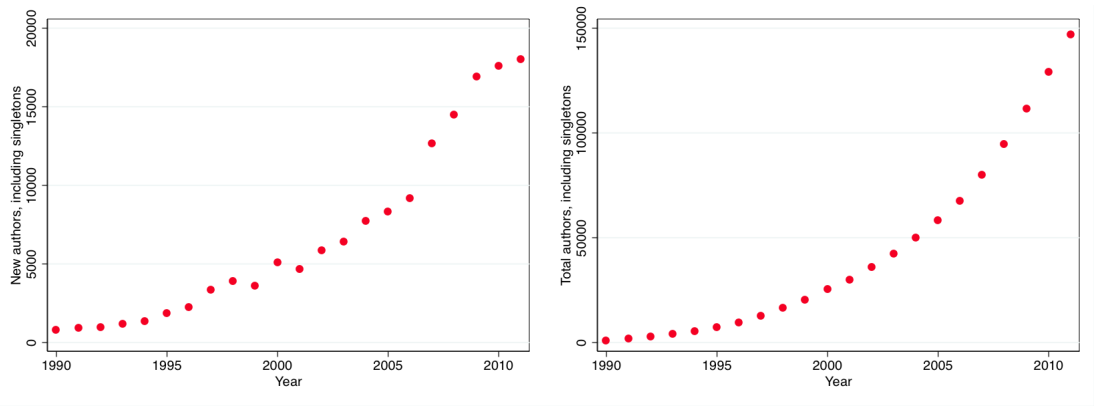


Figure 3. Including singletons, number of new authors per year (left) and cumulative number of authors (right).

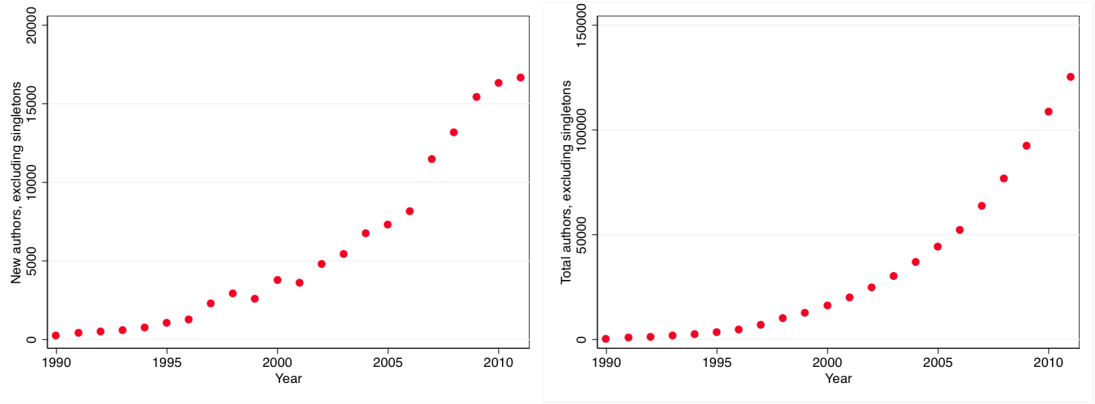


Figure 4. Excluding singletons, number of new authors per year (left) and cumulative number of authors (right).

Each year, there have been more and more individuals publishing about the Internet (either as individuals or with one or more co-authors). Beyond seeing that writing about the Internet has experienced a steady rise over the past two decades, there is not much more significance to this. Also note that writing across all fields has generally been increasing yearly (Larsen & von Ins, 2010). Although the null model of cold fusion shows that network growth is not an inevitability, it is hard to give a substantive interpretation about the significance of this amount of growth without a more directly comparable null model (relating to social science, and found by topic search).

The number of authors in the LCC (figure 5) is a more meaningful measure because it relates to a single extended community. It is of special interest if it departs from the growth of the overall network of all components.

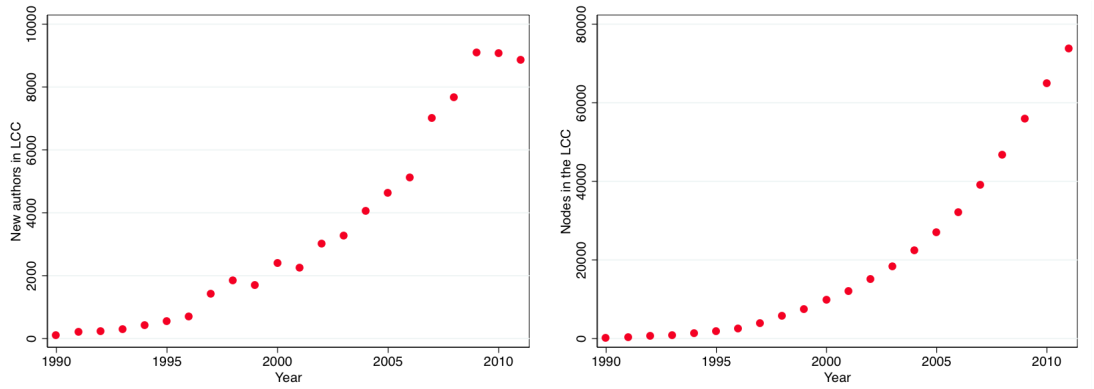


Figure 5. Number of new authors per year in the largest connected component (left) and cumulative number of authors in largest connected component (right).

One difference from figs. 3 and 4 is that in the past two years, the number of new authors per year in the LCC appears to be levelling off (fig. 5, left), although this is too short an interval from which to extrapolate (note the dips in 1999 and 2001 that did not turn into a trend). However, it might merit watching; perhaps the boundaries of the community of Internet research have stopped accelerating.

4.2. *Largest Connected Component*

In 2011, the largest connected component not only accounts for a majority of the network (50.20% of all nodes, and 58.90% of all co-authoring nodes), but it dwarfs the next-largest components (fig. 6, left): the LCC is 73,736 nodes, while the next-largest components is of size 78! While there is undoubtedly co-authorship structure in the network that is left out of this LCC, so long as it is dwarfed by the structure captured by the LCC, it is acceptable to look only at the LCC. As is often the case in various network metrics, there is a power law distribution for the component sizes (fig. 6).

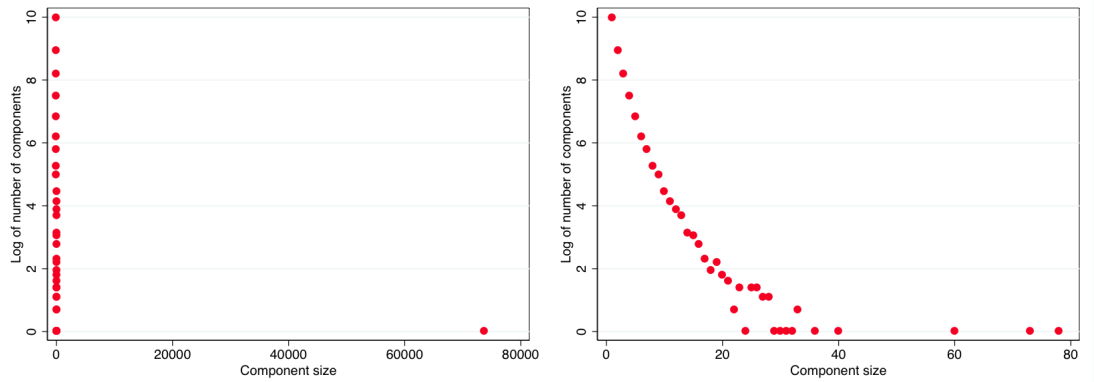


Figure 6. Number of components of each size on a logarithmic scale, including the LCC (left) to show the extent to which it dwarfs all other components and excluding the LCC (right) to show the variance in the sizes of the other components.

While we can see that the size of the LCC grows along with the network, we can see that it grows relatively more quickly than the network, as it accounts for a high fraction of the nodes in the network as time goes on (fig. 7).

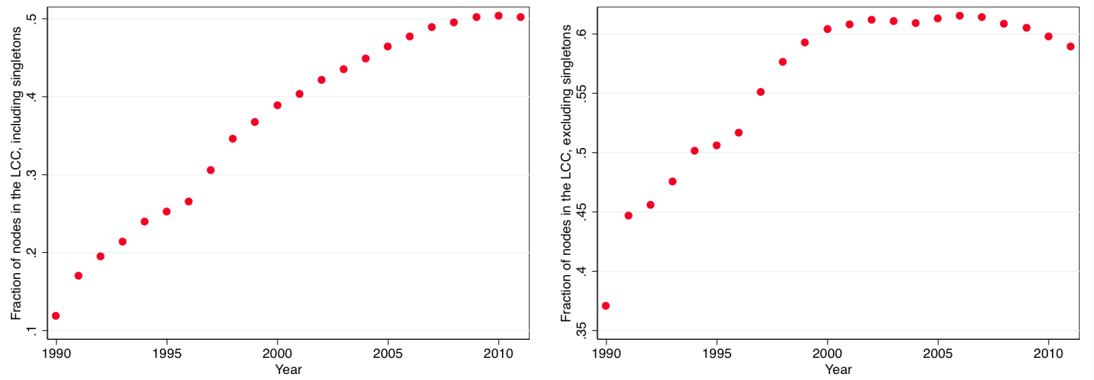


Figure 7. The fraction of the network represented by the LCC, including singletons (left) and excluding singletons (right).

Here we see nuance emerging from distinguishing between the network including singletons and one excluding singletons: if singletons are included, there is almost no levelling off of the fraction of nodes in the LCC, but excluding singletons, the LCC both starts as a larger piece of the network and then stops growing as a fraction around 2000, in fact dropping as time goes on. Given that we know the network has been growing overall, this means that since 2000, the number of authors entering the network in disconnected components of two or more authors has roughly stabilised proportional to the number of new authors added to the giant component per year. And in the past three or so years, the number of new authors added to the giant component has been growing faster than the co-authors entering the network outside the giant component.

Also note that these graphs look very different from the pattern shown below (fig. 8) in the fraction of edges, rather than the fraction of nodes, in the LCC over time. The amount of linkage, as captured by the fraction of edges in the LCC, is a different measure from the amount of coverage, as captured by the fraction of nodes in the LCC. Linkage tells about coherence, while coverage only provides an overview.

4.3. Coherence

As discussed above, I define "coherence" as the fraction of edges in the largest connected component, arguing based on the findings of Bettencourt et al. that we should interpret this fraction as representing the coherence of the network. The growth of the proportion of nodes in the LCC represents a growth in an interconnected community. And, identical to the findings of Bettencourt et al., Internet literature sees a strong growth in coherence over time (fig. 8).

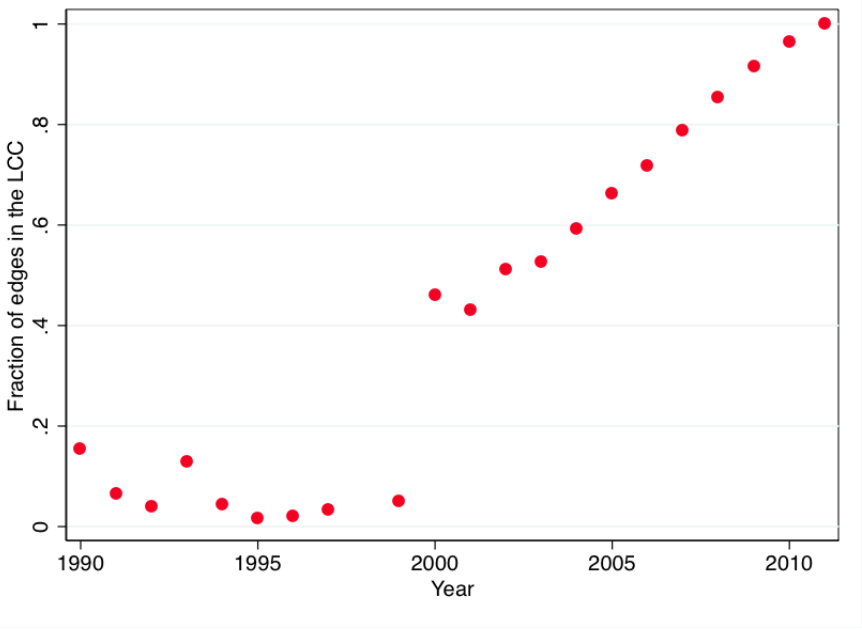


Figure 8. Fraction of edges in the largest connected component, which I call 'coherence', over time. Note the conspicuous gap at 2000, where the coherence jumps nearly tenfold from .0491 to .4600.

This is the key graph from which I make my argument that I have quantitatively identified the emergence of field of Internet studies. The pattern shown here is exactly the one Bettencourt et al. found to characterise successfully emergent fields, and a pattern that is not universal (as shown by the null model of cold fusion). Not only does this show a marked rise in coherence, but also there is specific structure—the intriguing behaviour around the critical time of 2000—that supports it having substantive meaning.[29]

The jump is consistent with the finding of Bettencourt et al. that such spikes in the fraction of edges in the LCC over time correspond to momentous moments in the field (fig. 2a). Cross-referencing the spike in coherence at 2000 to what we know about the development of the field

[29] I have made every effort to verify that this jump is real, and not just an artefact of the data: there is nothing in my data collection or processing that could produce this as a by-product. As also discussed above, there are no changes in the WOS database, no inconsistencies across time of its indexing system that would explain the result as an artefact. And other, independently gathered formed of data (such as the graph of JASIST articles about the Internet, fig. 1) also show distinctive behaviour at 2000.

of Internet studies around that time, I conclude that the spike is a network signature left by a deliberate push towards the social study of the Internet, and that this network signature shows that the push was successful.

Then, the steady pattern observable after the 2000 jump suggests that there were no further critical points after 2000, that the community of Internet researchers grew steadily upon the foundations established around 2000. This supports a view that, despite some activity around the middle of the decade and in recent years, it was really the initial push that gave the field its momentum and established a seed community that expanded.

4.4. Consolidation

Going beyond the approach of Bettencourt et al., I push further to examine the manner in which the network of Internet literature co-authorship became coherent.

As introduced in my methods discussion, I propose a new analytical technique where the edges of a later time are projected onto the nodes of an earlier time. The corresponding rise in coherence is an indication of how frequently edges in future years connect incumbent nodes (as of the given year) to incumbent nodes, rather than connect incumbent nodes to new nodes or new nodes to new nodes. A strength, and a weakness, of this approach is that it is defined in terms of hindsight; it is a strength in that it makes use of the opportunity of developing retrospective understanding, but it is a weakness in that it could not be adapted to have predictive power, and in that the findings would potentially change if repeated in future years when more data have been generated. To address the weakness, I will look not only at the increase in coherence that comes from projecting the edges of 2011 back on to nodes of a previous time, but the nodes of every year up until 2011 as well. Doing this involved organising information into a 21 x 21 upper triangular matrix (the upper left quadrant of which is shown in table 1 to illustrate).

Table 1. This is given to illustrate the data structure. The non-projection coherence scores for the given year are the diagonal entries (i.e., the nodes and edges of the same year will refer to the actual network for the year).

Nodes of: \ Edges of:	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
1990	0.1548	0.1548	0.1548	0.1765	0.1765	0.1765	0.1765	0.1765	0.1765	0.1765	0.1765
1991		0.0656	0.0656	0.0752	0.0749	0.0774	0.0900	0.0900	0.0897	0.0897	0.2215
1992			0.0395	0.0445	0.0444	0.0459	0.0542	0.0541	0.0557	0.0556	0.1402
1993				0.1282	0.1279	0.1275	0.0355	0.0353	0.0353	0.0351	0.0864
1994					0.0430	0.0429	0.0222	0.0539	0.0656	0.0776	0.1584
1995						0.0164	0.0202	0.0201	0.0465	0.0504	0.1267
1996							0.0197	0.0196	0.0398	0.0424	0.1052
1997								0.0321	0.0415	0.0550	0.1112
1998									0.0453	0.0564	0.1131
1999										0.0491	0.1121
2000											0.4598

Rather than generating 22 graphs, I will use this matrix to construct a 3d plot (fig. 10), but as it is difficult to understand the 3d plot without being able to dynamically rotate it,[30] I will provide six 2d graphs that are cross-sections for the 3d plot (fig 10).

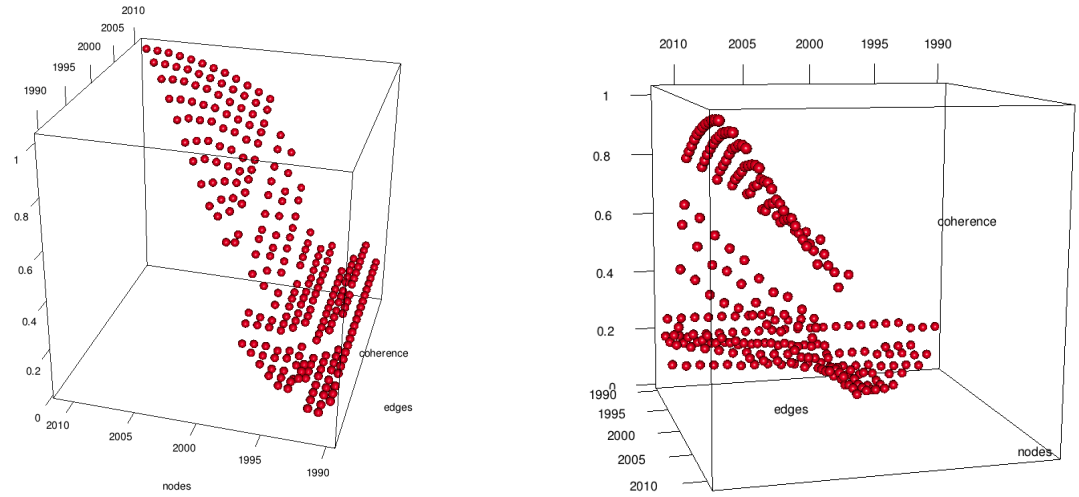


Figure 9. Two views of the 3d scatter plot of the projected coherence scores from 1990 to 2011. Cutting the diagram along the 'edges' axis produces the time slices shown in fig. 10 below.

[30] For the reader interested in dynamically exploring this graph, I have a put a script up in my Github repository (https://github.com/559755/Scripts-for-MSc-Thesis--Networks-of-collaboration-and-field-emergence-in--Internet-Studies--) that generates this graph (the data is included in the script). Running it requires R, as well as an installation of the R library "rgl" (http://rgl.neoscientists.org).

The 3d plots only occupying one diagonal half of the graph space comes from the matrix being upper triangular (the matrix is like the 'floor' of the graph, with a point above each cell at a height representing the coherence score within the cell). But visible is a smooth 'sheet' at the top of the graph, where coherence emerges across several time slices, and that grows higher as edges are projected from further in the future—a sign that the network underwent more consolidation as time went on. At the 'foot' of the graph, towards the nodes=1990 plane, the data become too choppy (an artefact of its sparseness at earlier years) to meaningfully interpret. In the middle, there is a precipitous drop at the nodes=2000 plane, where the projected edges show a drop in coherence across all time slices. This drop is more visible in time slices of the 3d plot (fig. 10).

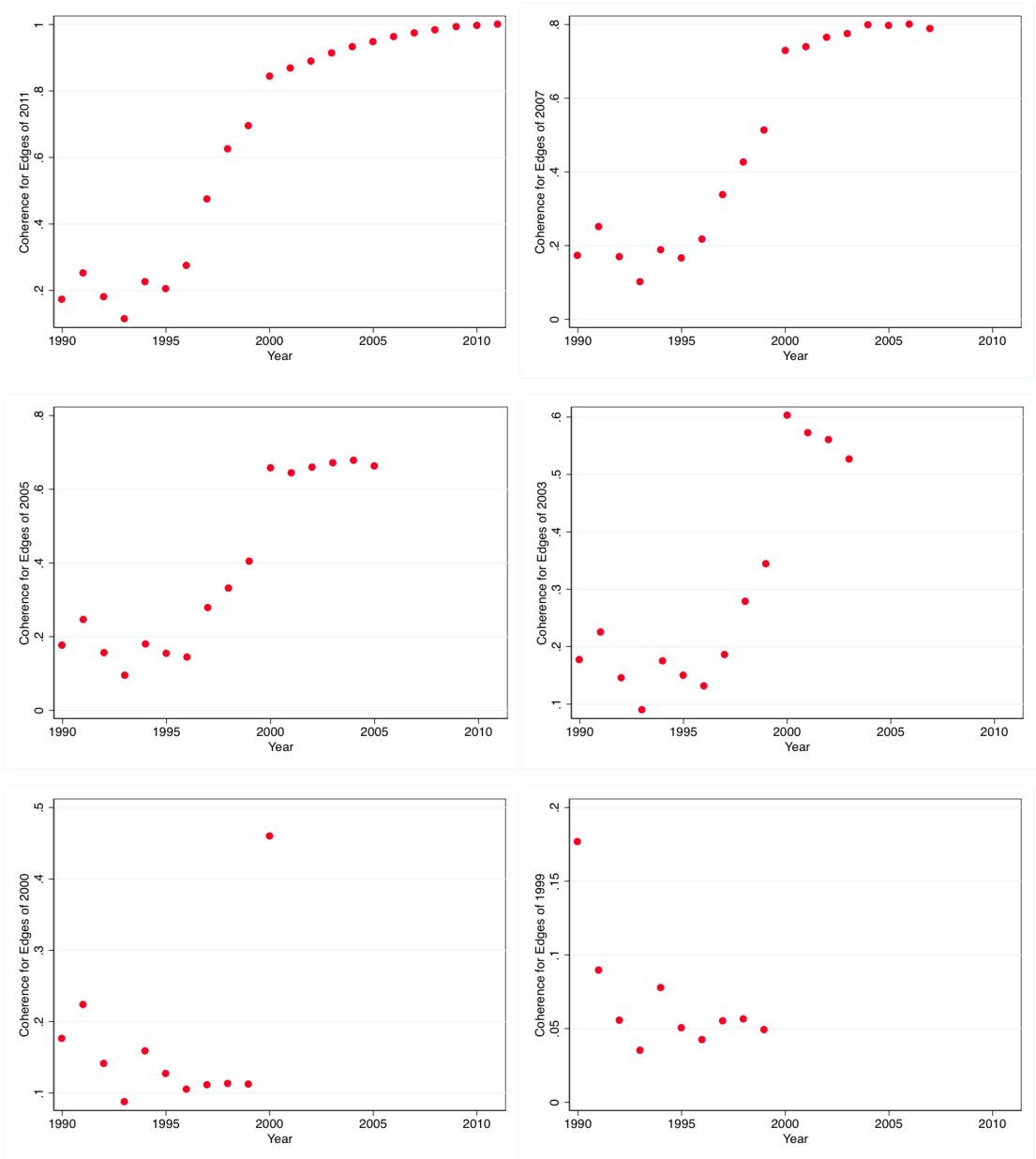


Figure 10. Coherence with the projected edges of 2011 (upper left), 2007 (upper right), 2005 (centre left), 2003 (centre right), 2000 (lower left), and 1999 (lower right). These may be understood as cross-sections along the 'edges' axis of the 3d plot in fig. 10. The irregular intervals (1999, 2000, 2003, 2005, 2007, 2011) are chosen to best illustrate variation.

Several interesting behaviours emerge. First, when projecting the edges of 2003, 2005, or 2011 back onto the late 1990s, there is a growth in the coherence of the LCC that portends the dramatic rise in coherence in 2000 that is not visible in the projected edges of 1999 or 2000 (nor the basic graph of coherence of figure 9). While we should ignore the coherence scores for the first few years across all graphs as messiness resulting from a small amount of data, in the graphs of the projected edges of 1999 and 2000, after 1995 we can see that the coherence of the network does not appear to be growing much. The change in pattern across the different projections shows that authors who were present before 2000 went on to author together after 2000. Second, the smooth rise of coherence after 2000 with the projected edges of 2011 is not present in other graphs. The projected edges of 2007 show a slight wavering, as do the projected edges of 2005. With the projected edges of 2003, there is in fact a decline in coherence! This indicates that for a brief period of time after the critical year of 2000, the network was going through a minor phase of spread, rather than consolidation. However, consolidation stabilised throughout the decade, resulting in a trend of smoothly rising levels of new collaborations between authors who had previously published on the Internet.

5. Discussion

5.1. Emergence of Internet studies

With this analysis, I provide quantitative evidence for a time at which we can say Internet studies emerged as a coherent entity. There are numerous lines of analysis supporting what Bettencourt et al. call a 'topological transition' happening in 2000: the stabilisation of the fraction of nodes in the LCC, the drastic jump in coherence at 2000, and the edge projection analysis showing that, for the projected edges of any year, a marked shift took place at 2000. Further support comes from the proportion of articles about the Internet in JASIST (table 1), which is a different data set but also shows a distinctive jump at 2000. While I do not know the factors behind this emergence (millennial fervour? Concerns about Y2K? The efforts of those in AoIR or newly established research centres? And do publishing delays mean the critical events would have happened in 1999 or 1998?), in this thesis I seek neither to address the "why" or "how" questions, nor to explore the nature of emergent community. My goal is to provide strong evidence-based answers to the "what" and "when" questions about the emergence of Internet studies.

5.2. Consolidation vs. Spread

The novel analysis I propose here has yielded some interesting results. Until 2000, the network was growing through spread, as most links formed between authors new to the network, and the projected edges from later networks yielded very low levels of coherence. A marked shift took place just before 2000, where the network began consolidating as a higher proportion of links began forming between authors who had already published about Internet studies. By 2000, the field had stabilised into a highly coherent entity. For a few years after, there was a short period of more spread, as the coherence for the projected edges of 2003 dropped for the nodes of 2001 and 2002. But soon after, the field returned to consolidation, at which point it finds itself today.

6. Conclusion

6.1. Field emergence

The first contribution I have made is towards the understanding of field emergence. Seeing that the longitudinal evolution of a cross-disciplinary co-authorship network has signatures of field emergence encoded in its structure adds to understandings of such emergence. This analysis could be extended to other emergent cross-disciplinary topics,[31] to see if co-authorship networks have evolved towards a great enough coherence to call the topic a field.

While the fraction of edges in the LCC in a co-authorship network is just a network signature, it links directly to the underlying dynamics of communication, collaboration, and relationship building, which are what define academic practice in a more grounded way than epistemologies, theories, or methodologies. The founding of journals, research centres, and professional bodies had the effect of bringing together individuals, resulting not just in co-authored papers, but in the construction of a giant connected component that links authors across the field. By giving the name "coherence" to the measure of the fraction of edges in the LCC, I embed this theoretical understanding in a new network metric that can contribute to future studies of field emergence.

[31] Two candidates I have recently come across are 'food studies' and 'memory studies'. This is not the place to justify that these might be emergent fields; I mention them only as examples.

6.2. Edge projection

The second contribution I make is new analytical technique of projecting later edges onto earlier nodes and measuring the change in coherence, and using it to extract substantive insights from longitudinal network analysis. If the coherence rises, it means that the edges formed in subsequent years form between incumbent nodes, meaning the network is undergoing consolidation where existing nodes link to one another. If the coherence falls, it means the network is spreading out, that new edges are being with new nodes. While not yet a formal statistical test, edge projection applies some of the insights behind network statistics in using longitudinal analysis to compare the network against itself (see Appendix B). And, compared to actor-based models of network statistics, my approach is not computationally intensive and hence can be applied easily to large networks.

6.3. Future steps

In order to make wider use of analysis of co-authorship networks based around topics rather than already well-defined fields of research, we will need better null models. The patterns of cold fusion found by Bettencourt et al. are invaluable for showing that common patterns are not just mathematical artefacts of the growth of network size, that they are not something inevitable and hence something we can substantively interpret. Still, a new null model is needed to cover *exploratory* analysis of the type I do here. For WOS records, this would involve randomly sampling articles in the database, and seeing the number of samples needed for a giant component to emerge or for the LCC to grow in a way that resembles the coherence of fields. At a certain (extremely large) size, I would expect a rising coherence score even with random sampling, because in some sense all of academia is one giant field: the field of academia. By this I mean that academia is a coherent entity in its deliberate social and institutional organisation, so at a certain size, it should show the same properties of coherence as do its subgraphs.

Another way I would like to extend this analysis in a future study is to have it be less self-contained—to combine analysis of the co-authorship network with some sort of disciplinary coding of authors. This could either be by cross-referencing the results of semantic analysis (such as that of Peng et al.) with community detection within the co-authorship network, or randomly sampling from within communities and performing content analysis or carrying out a qualitative investigation. For example, a future study could take other works by authors in the LCC and either code for disciplines or do a semantic network analysis to see if authors in the LCC publish exclusively about the Internet—not publishing exclusively about the Internet would show the extend to which Internet studies may serve as equivalent to a disciplinary home

(for a research agenda if not for a set of theories, methodologies, and a community), and to what extent it remains a secondary interest that is subservient to disciplinary identity. This would have more meaning for actual field practitioners, beyond being informative at a theoretical and global level about the dynamics of field emergence. Or, going into deeper theoretical relevance, work on field emergence through co-authorship (this thesis, Bettencourt et al.) could be cross-referenced with work on looking at field emergence through information networks (e.g., Chen, Chen, Horowitz, Hou, Liu, & Pellegrino, 2009). Such approaches would have to deal with the difficulties of analysing multiple overlapping networks, an area that is still developing mathematically (Mucha et al., 2010), but because the individual types of network analyses within bibliometrics have strong theoretical and empirical backing, multiplex bibliometry might be an ideal testing ground for developing the tools of multiplex network analysis.

6.4. Field self-knowledge

Beyond the omphaloskeptical exercise of having a precise time from which to date the 'beginning' of Internet studies, being able to cross-reference network signatures with the field meta-literature provides an empirical grounding for the self-knowledge of the field. Furthermore, while this is a purely descriptive thesis and does not comment on the normative questions about what the institutional status of Internet studies *should* be, it can help inform those debates.

For example, in her advice to the field of Internet studies in 2005, Baron (p. 271) suggests that for those who want to establish it as a discipline, to "to retain our current interdisciplinary standing for at least a few more years" in order to "endear ourselves to university administrators, who are generally disinclined to authorize new expenses in the current economic market", to "play an important collegial role by helping to revitalize some disciplines (sociology and linguistics come to mind) with our Internet research initiatives and by contributing to an important social agenda (especially with our colleagues in political science and sociology) through e-government initiatives", and to "manoeuvre into place a new generation of Internet researchers (with tenure) who (along with the earlier-tenured among us) can prepare for the next stage in the plan: structural independence." Revisiting her claim, it would appear that the rate of growth of a community of Internet researchers has stabilised, and there is a critical mass. Thus, if (but only if!) Internet scholars are dissatisfied in their current disciplinary homes (Ibid.), my analysis shows that now is as good a time as any to begin making a coordinated effort to establish an independent academic entity for Internet studies through,

along Baron's advice, "procuring funding for physical real estate", "establishing a public presence", and "developing a professional organization [the AoIR] that is recognized as setting the 'gold standard' for expertise on Internet matters".

On the other hand, following Sterne's (2005, p. 249) argument that Internet studies is already successfully reproducing itself, and hence "the symbolic benefits of becoming a discipline are relatively limited, and such a move would also have significant intellectual costs." Shrum (2005) similarly argues that the lack of a disciplinary nature of Internet studies is an intellectual asset. My analysis also gives empirical grounding and quantification to the claim that Internet studies is reproducing itself, and being successful without being a field. Thus, my evidence does not support the argument of either people desirous to establish a discipline of Internet studies, nor those who argue against it, but informs both about the state of the field. Whatever direction the social science of the Internet takes, having taken stock of its current trajectory through this thesis will hopefully help scholars in the field more effectively play a deliberate role in shaping its future.

Bibliography

Abbasi, A., Altmann, J., & Hossain, L. (2011) Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, *5*, 594-607.

Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A, 311*, 590–614. http://arxiv.org/abs/cond-mat/0104162

Bar-Ilan, J., Levene, M., & Lin, A. (2007). Some measures for comparing citation databases. *Journal of Informetrics*, *1*(1), 26–34

Baron, N. S. (2005). Who wants to be a discipline? *The Information Society*, *21*(4), 269–271.

Baym, N. K. (2005). Introduction: Internet research as it isn't, is, could be, and should be. *The Information Society*, *21*(4), 229–232.

Bellanca, L. (2009). Measuring interdisciplinary research: analysis of co-authorship for research staff at the University of York. *Bioscience Horizons*, *2*(2), 99-112.

Bettencourt, L. M. A., Kaiser, D. I., & Kaur, J. (2009) Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics*, *3*(3), 210-221. http://web.mit.edu/dikaiser/www/BKK.Topological.pdf

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*, *2008*(10), 1-12.

Börner, K., & Scharnhorst, A. (2009). Visual conceptualizations and models of science. Journal of Informetrics, 3(3), 161–172.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, *3*(3), 191-209.

Consalvo, M., & Ess, C. (Eds.). (2011). *The handbook of Internet studies*. Wiley-Blackwell.

Ding, Y. (2011a). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics*, *5*, 187-201.

Ding, Y. (2011b). Community detection: Topological vs. topical. *Journal of Informetrics*, *5*, 498-514.

Dutton, W. H. (2013, forthcoming). Internet studies. In W. H. Dutton (Ed.), *The Oxford handbook of Internet studies*. Oxford: Oxford University Press.

Dutton, W. H. (Ed.). (2013, forthcoming). *The Oxford handbook of Internet studies*. Oxford: Oxford University Press.

Feigenbaum, J., Hargittai, E., & O'Rourke, J. (1994, September). Expanding the pipeline: CRAW Database aids academic recruiters. *Computing Research News*.

Franceschet, M., & Constantini, A. (2010). The effect of scholar collaboration on impact and quality of academic papers. *Journal of Informetrics*, *4*, 540-553.

Gehrke, J., Ginsparg, P., & Kleinberg, J. (2003). Overview of the 2003 KDD cup. *Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, August 2003. http://www.cs.cornell.edu/home/kleinber/kddcup2003.pdf

González-Bailón, S., Borge-Holthoefer, J., Rivero, A., & Moreno, Y. (2011). The dynamics of protest recruitment through an online network. *Scientific Reports*, *1*(197).

Heersmink, R., van den Hoven, J., van Eck, N. J. P., & van den Berg, J. (2011). Bibliometric mapping of computer and information ethics. *Ethics and Information Technology*, *13*(3), 241-249.

Hennemann, S., Rybski, D., & Liefner, I. (2012). The myth of global science collaboration: Collaboration patterns in epistemic communities. *Journal of Informetrics*, *6*, 217-225.

Hine, C. (2005). Internet research and the sociology of cyber-social-scientific knowledge. *The Information Society*, *21*(4), 239–248.

Huang, M., & Chang, Y. (2011). A study of interdisciplinarity in information science: Using direct citation and co-authorship analysis. *Journal of Information Science*, *37*(4), 369–378.

Hunsinger, J., Klastrup, L., & Allen, M. (Eds.). (2010). *International handbook of Internet research*. Springer.

Jin, E., Girvan, M., & Newman, M. E. J. (2001). Structure of growing social networks. *Physical Review E*, *64*, 046132.

Jones, S. (2005). Fizz in the field: Toward a basis for an emergent Internet studies. *The Information Society*, *21*(4), 233–237.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Lambiotte, R., & Panzarasa, P. (2009). Communities, knowledge creation, and information diffusion. *Journal of Informetrics*, *3*, 180-190.

Larsen, P. O., & von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, *84*(3), 575-603.

Leskovec, J. (2008). Dynamics of large networks. PhD Dissertation, Machine Learning Department, Carnegie Mellon University. Technical report CMU-ML-08-111. http://cs.stanford.edu/people/jure/pubs/thesis/jure-thesis.pdf

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: Densification laws, shrinking diameters and possible explanations. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. http://cs.stanford.edu/people/jure/pubs/powergrowth-kdd05.pdf

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, *1*(1), Article 2. http://cs.stanford.edu/people/jure/pubs/powergrowth-tkdd.pdf

Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M. W. (2008). Statistical properties of community structure in large social and information networks. *International World Wide Web Conference 2008*, April 21–25, 2008, Beijing, China. http://cs.stanford.edu/people/jure/pubs/ncp-www08.pdf

Leydesdorff, L., & Wagner, C. S. (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics*, *2*, 317-325.

Leydesdorff, L., Carley, S., & Rafols, I. (2012). Global maps of science based on the new Web-of-Science categories. *Scientometrics*.

Liu, D., Blenn, N., & Van Mieghem, P. (2012). Characterizing the structure of affiliation networks. *Procedia Computer Science*, *00*, 1-10.

Markham, A. N. (2005). Disciplining the future: A critical organisational analysis of Internet studies. *The Information Society*, *21*(4), 257–267.

Mazar, R. (2010a). Appendix A: Degree programs. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International handbook of Internet research* (pp. 523-548). Springer.

Mazar, R. (2010b). Appendix B: Major research centers and programs. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International handbook of Internet research* (pp. 549-604). Springer.

McLemee, S. (2001, March 30). Internet studies 1.0: A discipline is born. *The Chronicle of Higher Education*.

Monberg, J. (2005). Science and technology studies approaches to Internet research. *The Information Society*, *21*(4), 281-284.

Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., & Onnela, J. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, *328*(5980), 876-878.

Newman, M. E. J. (2000). The structure of scientific collaboration networks. arXiv.org. http://arxiv.org/abs/cond-mat/0007214

Newman, M. E. J. (2001a). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404-409. http://www.cis.upenn.edu/~mkearns/teaching/NetworkedLife/newman-sci.pdf

Newman, M. E. J. (2001b). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, *64*, 016131. http://arxiv.org/abs/cond-mat/0011144

Newman, M. E. J. (2001c). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, *64*, 016132. http://arxiv.org/abs/cond-mat/0011144

Newman, M. E. J. (2003). The structure and function of complex networks. arXiv.org. http://www-personal.umich.edu/~mejn/courses/2004/cscs535/review.pdf, http://arxiv.org/abs/cond-mat/0303516

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, *101*(suppl. 1), 5200-5205. http://www.pnas.org/content/101/suppl.1/5200.long

Newman, M. E. J. (2006a, June). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577-8582. http://www.pnas.org/content/103/23/8577.abstract, http://arxiv.org/abs/physics/0602124

Newman, M. E. J. (2006b, September). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, *74*, 036104. http://arxiv.org/abs/physics/0605087

Newman, M. E. J. (2006c, May 4). Community Centrality. [Online supplement to Newman 2006b]. http://www-personal.umich.edu/~mejn/centrality/

Newman, M. E. J. (2010). Networks: An introduction. Oxford: Oxford University Press.

Newman, M. E. J. (2012, January). Communities, modules and large-scale structure in networks. *Nature Physics*, *8*, 25-31.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*, 026113. http://arxiv.org/abs/cond-mat/0308217

Obermeier, U., & Brauckmann, H. (2010). Interdisciplinary patterns of a university: Investigating collaboration using co-publication network analysis. *Collnet Journal of Scientometrics and Information Management*, *4*(1), 29-40. http://arxiv.org/abs/1003.4131.

Peng, T. Q., Zhang, L., Zhong, Z. J., & Zhu, J. J. H. (2013, forthcoming). Mapping the landscape of Internet studies: Text mining of social science journal articles 2000-2009. *New Media & Society*.

Persson, O., Glänzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, *60*(3), 421-432.

Qin, J., Lancaster, F. W., & Allen, B. (1997). Types and levels of collaboration in interdisciplinary research in the sciences. *Journal of the American Society for Information Science*, *48*(10), 893–916.

Ribes, D., & Bowker, G. C. (2008). Organizing for multidisciplinary collaboration: The case of the geosciences network. In G. M. Olson, A. Zimmerman & N. Bos (Eds.), *Scientific collaboration on the Internet* (pp. 311-330). Cambridge, MA: The MIT Press.

Rice, R. E. (2005). New media/Internet research topics of the Association of Internet Researchers. *The Information Society*, *21*(4), 285-299.

Rodriguez, M. A., & Pepe, A. (2008) On the relationship between the structural and socioacademic communities of a coauthorship network. *Journal of Informetrics*, *2*, 195-201.

Rombach, M. P., Porter, M. A., Fowler, J. H., & Mucha, P. J. (2012). Core-periphery structure in networks. http://arxiv.org/abs/1202.2684.

Saavedra, S., Reed-Tsochas, F., & Uzzi, B. (2008). Asymmetric disassembly and robustness in declining networks. Proceedings of the National Academy of Sciences, 105(43), 16466–16471.

Shrum, W. (2005). Internet indiscipline: Two approaches to making a field. *The Information Society*, *21*(4), 273-275.

Silver, D. (2004). Internet/cyberculture/digital culture/new media/fill-in-the-blank studies. *New Media & Society*, *6*(1), 55-64.

Snijders, T. A. B., van de Bunt, G. G., & Steglich, C. E. G. (2010) Introduction to stochastic actor-based models for network dynamics. *Social Networks*, *32*, 44-60.

Sterne, J. (2005). Digital media and disciplinarily. *The Information Society*, *21*(4), 249-256.

Van Eck, N. J. P., & Waltman, L. (2007). Bibliometric mapping of the computational intelligence field. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *15*(5), 625-645.

Wellman, B. (2004). The three ages of internet studies: ten, five and zero years ago. *New Media & Society*, *6*(1), 123-129.

Wellman, B., Frank, O., Espinoza, V., Lundquist, S., & Wilson, C. (1991). Integrating individual, relational and structural analysis. *Social Networks*, *13*, 223-50.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007, May 18). The increasing dominance of teams in production of knowledge. *Science*, *316*, 1036-1039.

Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cogitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*, *63*(7), 1313–1326.

Appendix A: Data collection details

I depart from Peng et al.'s (2013, forthcoming) search terms in several ways. First, I start my data collection at 1990. I originally started with 2000 as did Peng et al., but I found that this floor was insufficient for the longitudinal analysis I wanted to conduct modelled after Bettencourt et al. (who look at a minimum of 25 years). While the WOS index starts from 1926, and my query from 1926-1989 yields 10,167 hits, most seem to be proper names including "Web", such as Weber. Without "Web", there are only 3,864 hits, but either excluding "Web" or filtering individual proper names that include Web would have made my search query inconsistent with the latter query (and in later time periods, I did not want to exclude instances of people with a Web-containing surname publishing about the Internet). Instead, I chose 1990 a new floor, mostly because the World Wide Web was invented in 1990. I could have started at 1995, when Silver (2004) identifies Internet studies as starting, or with 1994, but I chose a few years before this to allow my network to have a sufficient 'base mass' for the emergence of a LCC, as well as to include some years without activity in order to have a smooth tail for various time graphs. Another work supporting the choice of 1990 is Rice's (2005) analysis of the instances of the term "Internet" in five disciplinary databases from 1985-2003; he finds a zero percent mention of "Internet" before 1990, and only the Library and information science database had a greater than zero percent mention of the Internet before 1993 (at which time the databases Business Source Premier, SocioFile, Comm Abstracts, and MedLine increased to having nonzero percentage representation; Rice, 2005, p. 286).

Second, while Peng et al. search only for "cyber space" OR "cyberspace", I use the general prefix "cyber-". In my explorations, I found only two terms that begin with cyber that do not explicitly related to the Internet: cybernetics and cyberpunk, both of which have negligible representation (1,181 and 139 hits, respectively). Second, contrary to Peng et al., I decided not to do post-collection filtering for "web". I found even in Peng et al.'s data, some articles about "food webs" that are not relevant to Internet studies. However, I found that filtering out "food web", or "web of science" as did Peng et al. (through the Boolean 'NOT' operator in the WOS search query), again yielded a negligible difference. I made a decision to keep the query as simple as possible, and not to filter out all possible false positives, as the process of finding false positives was largely ad-hoc and the amount filtered out for each query turned out consistently to be negligible. Again, I was guided by the understanding that I could safely 'overshoot' and thus should try to be as liberal as possible with data collection, because all false positives would be filtered out in the process of looking only at the LCC.

After some experimentation, I added a number of terms that yielded significant amounts of results (in the order of a few thousands). This included the names of various Internet-specific companies, sites, or organisations (Google, Facebook, Twitter, MySpace, YouTube, eBay, Wiki)[32]; any articles about any of these will most definitely pertain to the Internet. Another class is the various "e-" prefixed terms. Unfortunately, I had to manually generate a list of these terms; especially because the e- prefix is often not hyphenated, there is no systematic way to distinguish e- prefixed Internet-related terms from English words that begin with "e". For comparison, searching only for TS=(Internet OR cyber* OR online OR on-line) along the same parameters yielded 59,993 hits.

While at first I planned only to collect data from 2000-2009, and in fact to try and replicate Peng et al.'s exact data set, I quickly realised that (as explained above) I would gain a much richer data set by overshooting, and so long as I was going beyond the bounds of Peng et al.'s query, I might as well include all hits until the last complete year (i.e., 2011). By the same token, I included Conference Proceedings Citation Index– Social Sciences and Humanities (CPCI-SSH); as it begins in 1990, which corresponds to the floor of my data set.

The Book Citation Index– Social Sciences and Humanities (BKCI-SSH) proved to be more challenging to treat. The key problem is that it begins in 2005, three-quarters of the way through the time interval of my data set. There are no exact figures for how many books are in the index per year, [33] and no way to do a 'blank' search within the WOS to find this out directly. Searching for the 100 most common words in the English language[34] in only BKCI-SSH yielded 243,358 total results, with no hits prior to 2000 (table 2).

[32] My choice of what to include or exclude was based on listings such as that on http://www.alexa.com/topsites, but also on my own knowledge of what companies have received academic attention (e.g., MySpace is no longer important, but was earlier on in the decade and hence would be appropriate to include). I also excluded names based on experimentation. For example, I excluded "Amazon" because it returned articles about the South American biome. Or, "LinkedIn" did not yield an appreciable number of results so I excluded it. There are likely many more names I could have included, but this seemed a good selection.

[33] The WOS claims it was 30,000 initially, and has grown by 10,000 per year. http://wokinfo.com/media/pdf/bkci_fs_en.pdf. However, my analysis finds more than double what would be expected from 30,000 + (2011-2005)*10,000, so the WOS figure is not reliable.

[34] The query "TS=(the OR be OR to OR of OR "and" OR a OR in OR that OR have OR I OR it OR for OR "not" OR on OR with OR he OR as OR you OR do OR at OR this OR but OR his OR by OR from OR they OR we OR say OR her OR she OR "or" OR an OR will OR my OR one OR all OR would OR there OR their OR what OR so OR up OR out OR if OR about OR who OR get OR which OR go OR me OR when OR make OR can OR like OR time OR no OR just OR him OR know OR take OR people OR into OR year OR your OR good OR some OR could OR them OR see OR other OR than OR then OR now OR look OR only OR come OR its OR over OR think OR also OR back OR after OR use OR two OR how OR our OR work OR first OR well OR way OR even OR new OR want OR because OR any OR these OR give OR day OR most OR us)". The quotes around 'and', 'or', and 'not' are to distinguish them from Boolean operators. The list is from http://oxforddictionaries.com/words/the-oec-facts-about-the-language. Note that an earlier, unsystematically chosen search I did, "TS=(the OR a OR of OR on OR in OR at OR for OR to OR by OR from OR "and")" yields 234,826 results, suggesting diminishing returns.

Table 2. WOS hits in BKCI-SSH from 1990 to 2011 for the hundred most common English words, used as a proxy for total articles. There are no results prior to 2000.

Year	Records
2000	53
2001	77
2002	250
2003	6,859
2004	8,697
2005	17,838
2006	27,021
2007	33,260
2008	33,711
2009	42,429
2010	32,865
2011	40,298
Total	243,358

While there are some records before the index starts in earnest in 2005, the introduction of BKCI-SSH is still a sudden change in the data structure. On the one hand, it is good to have uniformly flawed data, where all parts suffer from the same shortcomings, especially where the uniformity is over time for the purpose of a longitudinal analysis. On the other hand, I want to have as rich a data set as possible, and including books would certain be an opportunity to capture as many co-authorship connections as possible.

I ultimately decided to include books, as I found that searching across the SSCI, A&HCI, CPCI-SSH, and BKCI-SSH yielded 114,079 hits from 1990-2011, whereas excluding BKCI-SSH yielded the only marginally smaller amount of 110,352 (with no hits in 2000 or 2001, 2 in 2002, 97 in 2003, 87 in 2004, and about 600 hits per year from 2005-2011). I decided that the potential for biasing the longitudinal analysis was miniscule, and so—while the chance that leaving out BKCI-SSH would fragment the LCC would be equally miniscule—I would err on the side of pursuing the richer data set. Note also that my main findings will pertain to 2000, and there are hardly any entries in the BKCI-SSH for around that time to influence that result.

Appendix B: Rejected analyses

Use of databases other than the WOS

I considered using Google Scholar, but Bar-Ilan, Levene and Lin (2007, p. 29) found it to be so inconsistent that in order to compare it to other bibliographic databases required a small data set with extensive manual cleaning. Bettencourt et al. (2009, p. 213), while compiling multiple other databases, decided to avoid Google Scholar entirely because of its inconsistency. Perhaps the only self-contained alternative to the WOS is Elsevier's relatively new Scopus database, but the use of this is not as established in literature and I decided not to explore it.

Using the Book Authors (BA) and Book Editors (BE) fields along with AU

At first, I planned to include Book Editors (BE) and Book Authors (BA) along with authors. I reasoned, for an author to publish in an edited volume was a sufficiently meaningful link to be worth including in a collaboration network. However, it seemed to be that I would then have a problem with nomenclature: could I still call it a co-authorship network? 'Collaboration network' is a more loose designation, but the phrase is still associated with co-authorship networks and, for academic work, considered synonymous with them. Although I will have lost some information through the decision to omit consideration of the BE and BA fields, it will maintain consistency with the wider literature. Note also that these fields did not exist in previous versions of the WOS. Note also that generally, for books, the AU field will list the book authors. In my data, there were only four non-singleton cases where the record contained only BA with AU being blank. Taking into account these cases had no effect on the network (or rather, one of these caused an otherwise unconnected 3-node component to connect to the giant component, but this is negligible).

Centrality measures

I computed centrality measures, but found them to not be very interesting. For example, the author with the highest eigenvector centrality (removing edge weights, and leaving multiple lines) is P. McGuigan, who shows up in the database as a co-author only on two papers, both with dozens of co-authors. It seems very likely, then, that McGuigan's centrality is an artefact of the calculation method rather than something with substantive meaning. Specifically, it relates to a problem that measures such as centrality or clustering do not take into account the artificially inflated clustering that arise from affiliation networks (oh which co-authorship networks are an example).

As an example, if there are five co-authors on a paper, in a 2-mode network of articles and authors, there would be a single node with five connections to five other unconnected nodes. When this is projected into a 1-mode network of co-authorship, where an edge represents co-authorship on a paper, the five authors will form a dense, fully connected cluster (fig. 11).

Figure 11. A 2-mode affiliation network (left) and its 1-mode projection (right). The clustering coefficient goes from 0 (no clustering) to 1 (total clustering).

Correcting for the artefacts created by projecting affiliation networks is an active research area (Liu, Blenn, & Van Mieghem, 2012), but it is not yet one that has an established tradition of application.

Network statistics with actor-based modelling

I do not carry out statistical investigation (i.e., hypothesis testing with tests of significance) of my network structure or dynamics, as network statistics is a still-developing research area. The state of the art in this respect is the longitudinal, actor-based SIENA (Statistical Investigation for Empirical Network Analysis) model (Snijders, van de Bunt, G. G., & Steglich, 2010), which models the probability of a tie between any two nodes being created (or broken). This is also a model that takes into account more than just the LCC, because it models the probability of ties being formed where there currently are none.

Ideally, I would use the SIENA model for exploring the dynamics of network emergence. Unfortunately, its implementation is computationally intensive and can only handle networks of at most a thousand nodes, while my network is two orders of magnitude larger. Still, I take

inspiration from one of its central insights, that networks may be used for hypothesis testing by comparing the network against itself across different points in time. That is, the manner in which a network grows (or shrinks) is how to understand the dynamics behind the network.

Regression to predict co-authorship based on discipline

I would also have liked to run regressions on the network to see if working in the same field is a stronger predictor of co-authorship or not. However, this gets back to the difficulty of network statistics, and the difficulty of finding a null model. In this case, note that being in the same field would make authors more likely to publish together on *any* topic, not just on the Internet. Even an informal regression is difficult because of the incompleteness of authorial disciplinary affiliations. In WOS data, institutional affiliations are only listed for the first author, and even then, inconsistently. There are several other potentially relevant fields (see Table A), but none of them is sufficient either: PA is the correspondence address, but it lacks department names, and is substituted (without a mark) for the publisher's address when the author address is not available. And, it applies only to the first author. EM is email address, but again for only the first author, and it is often not helpful at identifying affiliation: only 31,820 records out of the 114,079 total had an email address with a '.edu' or '.ac' in it. While some .com addresses may well be from company research outfits like Bell Labs, it would require making a manual list to distinguish those from generic addresses like those at gmail.com. The WOS does provide subject codes, SC and a new classification scheme for WOS v5, WC (Leydesdorff, Carley, & Rafols, 2012), but these are not disciplinary identifications; and, at 222 and 225 categories respectively, and thus do not contain much more information that does a semantic analysis or keyword search. Lastly, PI is the publisher city, not the location of the author.

Table 3. Web of Science fields. Note that for most records, most of these fields are empty. Source: http://images.webofknowledge.com/WOKRS56B5/help/WOS/hs_wos_fieldtags.html

Code	Field
PT	Publication Type (J=Journal; B=Book; S=Series)
AU	Authors
BA	Book Authors
BE	Editors
GP	Book Group Authors
AF	Author Full Name
CA	Group Authors
TI	Document Title
SO	Publication Name
SE	Book Series Title
LA	Language

DT	Document Type
CT	Conference Title
CY	Conference Date
CL	Conference Location
SP	Conference Sponsors
HO	Conference Host
DE	Author Keywords
ID	Keywords Plus®
AB	Abstract
C1	Author Address
RP	Reprint Address
EM	E-mail Address
FU	Funding Agency and Grant Number
FX	Funding Text
CR	Cited References
NR	Cited Reference Count
TC	Web of Science Times Cited Count
Z9	Total Times Cited Count (WoS, BCI, and CSCD)
PU	Publisher
PI	Publisher City
PA	Publisher Address
SN	ISSN
BN	ISBN
J9	29-Character Source Abbreviation
JI	ISO Source Abbreviation
PD	Publication Date
PY	Year Published
VL	Volume
IS	Issue
PN	Part Number
SU	Supplement
SI	Special Issue
BP	Beginning Page
EP	Ending Page
AR	Article Number
DI	Digital Object Identifier (DOI)
PG	Page Count
WC	Web of Science Category
SC	Subject Category
GA	Document Delivery Number
UT	Unique Article Identifier

Existing research that looks into disciplinarity from the WOS relies on outside reference material to code for disciplinary coding (e.g., Bellanca, 2009; Obermeier & Brauckmann, 2010; Huang & Chang, 2011; Qin, Lancaster, & Allen, 1997). For Internet studies, there is no organising listing. And, unfortunately, there is no comprehensive, central listing of academics and their affiliations. There is CollabSeer, but it is a database mostly only of computer scientists. Then there is the new Microsoft Academic Search (MAS) engine

(http://academic.research.microsoft.com/), but it is still in beta. In some experimentation with this, I found that MAS is not yet very good at finding academics based only on last name and first initials, but using MAS as a cross-reference to find authorial affiliations may be feasible given enough effort.

Comparing communities to the clusters of Peng et al.

The analysis I would have most liked to do, but that was currently not feasible, is to see if I could detect communities and see how well they map on to the analysis of Peng et al. While I did explore doing this, I found that authors frequently authored papers across multiple of Peng et al.'s subject categories. Since the categories are categorical, there is no way to take an average, and in most cases of multiple categories, there was no modal category (e.g., there would be a single author publishing on three papers, each falling into a different category). It would be possible to assign multiple, simultaneous node attributes, but carrying this out would require a great deal of custom programming that would have been a thesis onto itself. Furthermore, Peng et al. only have codes for some 25 thousand articles, whereas I had 114,079 (or, excluding articles by only one author, I have 60,400 articles, and Peng et al. have 18,200), meaning Peng et al.'s classifications would be extremely sparse within my data set. And replicating Peng et al.'s methodology of two-stage cluster analysis for constructing co-occurrence networks would, again, be an entire thesis onto itself.

Then there is the problem of community detection. While community detection algorithms are applied to projected co-authorship networks (Newman, 2001a; 2012; Rodriguez & Pepe, 2008; Ding 2011a; 2011b), the state of the art Louvain method (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) still only detects discrete communities; it cannot detect overlapping communities, nor can it detect other potential structures such as core-periphery, or combinations of core-periphery structure and community structure (Rombach, Porter, Fowler, & Mucha, 2012).

I did apply the Louvain method through a NetworkX implementation: the partition calculated by the algorithm gave 208 communities, ranging in size from 3444 authors to 5 authors (fig. 12).

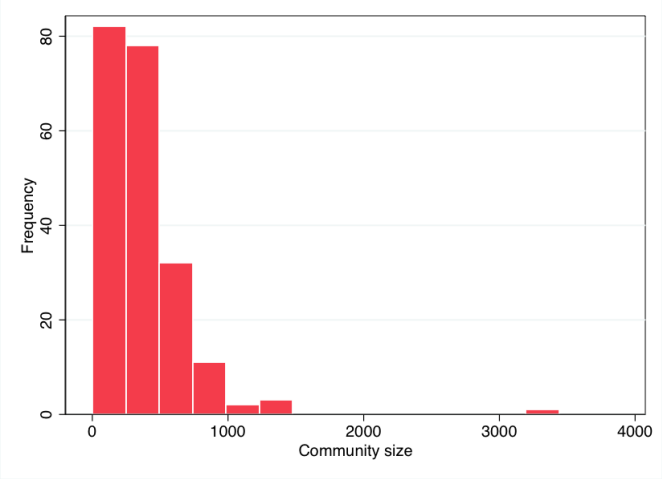


Figure 12. Histogram of size of communities detected by Louvain method.

Without having a methodology by which to look inside communities—such as cross-referencing with Peng et al., or cross-referencing with external disciplinary affiliation data, or even randomly sampling authors from within communities and manually finding out their disciplinary affiliation or main topics of study—knowing what is calculated by the Louvain method provides just lists of algorithmically clustered authors devoid of any meaning.

Using ready-made tools for WOS data analysis

There do exist ready-made tools for analysing WOS data. WOS provides its own citation analysis, as well as its own "HistCit" tool.[35] These allow some types of analyses, but not the network metrics explored here; this required me to construct networks from scratch. Then, despite being designed to accommodate WOS results, VOSviewer (http://www.vosviewer.com/) was unable to handle the volume of data I was processing with available computational power. VOSviewer also has been used for concept analysis (Heersmink, van den Hoven, van Eck, & van den Berg, 2011; van Eck & Waltman, 2007), but this requires the input of a separate (manually made) thesaurus file. Again, a combination of the results of Peng et al. and this approach might be fruitful, but as mentioned above, this would involve managing multiple node attributes.

Network visualisation

By the same token, although there is a large literature about visualisation (e.g., Börner & Scharnhorst, 2009), I found that my network was too large to visualise with the computational

[35] http://thomsonreuters.com/products_services/science/science_products/a-z/histcite/

power available. The only way to visualise the network would have been to condense it into communities, and graph the connections between communities. For example, I could have used the Louvain method for this, but without having a way to substantively label the detected communities, graphing them would have been meaningless.

Citation network analysis

I explored making a citation network, using the Cited References (CR) field (which contains all citations made by the given article in an abbreviated form, separated by semicolons); this was a problem because the citations are often quite inconsistent. While the WOS's internal system makes citation links through a unique WOS ID, the exported citations do not give this ID even when available. More annoyingly, citations will be (when consistent and regular) in the form "LNAME FM, YYYY, ABBREV J NAME, DOI;", a form in which the original records themselves are not given. I did make two citation networks; for one, I used only those records and citations for those records which had a DOI (about half of citations have a DOI, a document object identifier, a universal way to uniquely identify articles), using regular expressions to extract instances of DOIs from the CR field, and for the other, I wrote a script to rewrite primary records into the 'cited record' format, but I was unsure about the consistency of this result (and there is no documentation I could find about the exact rules by which WOS codes citations). Because of these difficulties, I put aside the citation networks and focused only on the far more consistent network of co-authorship.

Network decay

One weakness of co-authorship networks is that they represent a tie that was made at one point in time, but ties of collaboration do not stay indefinitely. After a time, it is no longer sociologically appropriate to maintain a tie. However, while ties are unambiguously formed through the process of co-authorship, there is no way of knowing when the ties break off, or how they 'decay'. Unfortunately, aside from giving a decay parameter to each tie, which would be rather arbitrary, there is no way with co-authorship networks to factor in decay. Network decay is an enormously important area of study, but the difficulties in addressing it mean it remains neglected (Saavedra, Reed-Tsochas, & Uzzi, 2008).