Mapping literary groups using graph theory and online reference works

Louis Goddard

May 13, 2016

Abstract

This paper explores the relevance of graph theory—and its practical offshoot, social network analysis—to macro-level literary history. It presents two primary techniques in a manner accessible to non-technical scholars in the humanities: one for extracting a social graph from online reference works such as *The Oxford Companion to Modern Poetry in English* and another for programatically detecting literary groups—also known as 'circles', 'schools', 'coteries' or 'movements'—within that graph. The efficacy of the *Walktrap* community detection algorithm is tested against the existing model of twentieth-century literary history, while the contours of that history are themselves further delineated through computational analysis. This double movement raises basic questions about the scope of the digital humanities and the status of data in the discipline. Specifically, it encourages a conception of data—and particularly metadata—not as a digital representation of pre-existing information but as substantially new information created in the process of digitisation.

The identification of groups of writers has been a staple of literary studies since its modern foundation in the late 19th century. Whether referred to as a 'circle', a 'school', a 'coterie' or a 'movement'—terms which of course carry differing implications—the literary group is, like the literary period, one of the foundational concepts upon which historical analysis is constructed. Again like the literary period, the literary group is a site of perpetual contestation, disputed both by scholars and, in the study of modern and contemporary literature, by the very writers who it attempts to subsume. A particularly stark example is the so-called 'Cambridge School', a category which is by now so well-established that it appears in reference works, yet of which membership has seemingly never been positively claimed by a single writer, even in retrospect.[1] In addition to problems of initial categorisation, the development of writers' careers threatens to draw them apart over time: while Edward Upward and Christopher Isherwood formed an extremely tight-knit coterie of two during their surrealistic 'Mortmere' phase of the 1930s, their very different work in subsequent decades—Isherwood's bourgeois realism and Upward's deeply political variety—would be difficult to group together at all.[2] These uncertainties are inherent in the structure of the concept and are not susceptible to empirical resolution: simply collecting more and better information on writers and their work will not allow future literary historians to produce a 'grand, unified theory' of literary groups. Nevertheless, it is reasonable to expect that the process of grouping itself might be further clarified, and that this second-order aim might be pursued in a broadly empirical fashion. This paper presents a programmatic method for extracting groupings from existing literary-historical data, using human-defined links between individual writers but without relying on human-defined groups.

[1] Nigel Wheale, 'Cambridge School, The', in *The Oxford Companion to Modern Poetry in English*, ed. by Ian Hamilton and Jeremy Noel-Tod (Oxford: Oxford University Press, 2013), <http://www.oxfordreference.com/view/10.1093/acref/9780199640256.001.0001/acref-9780199640256-e-1383/> [accessed 21 March 2016]. An example of such retrospective acknowledgement in relation to another group is Donald Davie's essay 'Remembering the Movement', an embarrassed account of his own involvement in the eponymous group anchored by Philip Larkin and Kingsley Amis (Donald Davie, 'Remembering the Movement', in *The Poet in the Imaginary Museum*, ed. by Barry Alpert [Manchester: Carcanet Press, 1977], pp. 72–75).

[2] For the former, see Christopher Isherwood and Edward Upward, *The Mortmere Stories* (London: Enitharmon Press, 1997).

The gradual shift of scholarly resources from print to screen has had profound implications for many aspects of the academy, perhaps the most obvious being accessibility: works previously held on physical library shelves are now available globally in digital form.[3] Less often considered are the processes by which digitisation creates new information, even when it appears simply to be duplicating what already exists. One area in which these processes are particularly visible is that of reference works such as encyclopaedias. Following the hypertext conventions of the web, and particularly of the free online encyclopaedia Wikipedia, digitised reference works now frequently include hyperlinks between related entries. Going beyond the 'see also'-type cross-references of their print equivalents, these links are embedded in specific parts of the entry text, and may link to other entries in ways which would be considered too tangential for a formal cross-reference. Importantly, these links are relatively trivial to extract and map—at least in comparison to print cross-references—and therefore offer a substantially new tool for the analysis of a reference work's contents.

In order to test the usefulness of hyperlinks in establishing literary groups, a reference work was selected which matched the author's own field of specialism: *The Oxford Companion to Modern Poetry in English* (*OCMP*). Originally published in 1994 as *The Oxford Companion to Twentieth-Century Poetry in English*, edited by Ian Hamilton, a second edition of this work was edited by Jeremy Noel-Tod and published in 2013, alongside a digital version on Oxford University Press's *Oxford Reference* website (oxfordreference.com). The book's link network was extracted by a relatively simple 'crawler' script written in the Python programming language, which loaded each entry in turn and recorded all links in the main text of the page, as well as the name and date of birth in the case of entries about individual poets; the relatively small number of entries not about poets was excluded from the data by hand.[4] Programmatically comparing the resulting link lists yielded a graph (in the mathematical sense), with poets as vertices (i.e. points or nodes) and links between their entries as edges (i.e. connections). Comparing the birthdates of each poet and rebasing the difference to a number between 0 and 1 (0 = 121 years, 1 = 0 years) generated a weight for each edge; in other words, the closer two poets were in age, the stronger their connection. Figure 1 shows the full graph, consisting of 1113 vertices and 2563 edges, plotted using the DrL force-directed layout algorithm with labels omitted for clarity. Like many force-directed layout algorithms, DrL simulates attractive and repulsive forces between vertices based on their connection (or non-connection) by edges; the resulting graph shows the system in a state of equilibrium. The width of the line representing each edge is determined by the weight of the connection.[5]

The full graph of the *OCMP*'s link network is unwieldy, to say the least. Even at the macro level, however, it offers useful information about the book's organisation. The *OCMP* appears to be organised around two densely connected cores of poets—the two large black zones in Figure 1—and a third, slightly less dense sub-core. These cores are connected by two relatively dense bridges, with a number of distinct spurs jutting out from the structure as a whole. Surrounding but unconnected to the central cluster is a fringe of singletons—poets unconnected to any other poet—and small clusters of between two and seven. These small clusters are the first clear candidates for recognition as literary groups in this analysis. The largest, for example, consists of six Nigerian poets and one Ugandan: Michael Echeruo, Richard Ntiru, Chimalum Nwankwo, Moses Tanure Ojaide, Christopher Okigbo, Remi Raji and Wole Soyinka. This structure reflects a network of influence centered on Okigbo (b. 1932), the oldest poet of the group; indeed, all members except Soyinka are connected directly to Okigbo. There are also sociological similarities: all except Ntiru and Okigbo himself have held academic positions in the United States, exerting a strong influence on the Anglo-American critical conception of African poetry in English. Yet they remain peripheral—a descriptive rather

[3] The issue of accessibility nevertheless remains a fraught one even in the digital arena, as the recent controversy over Sci-Hub, a free online repository of journal articles, made clear.

[4] Due consideration was given to the potentially negative effects of web crawling on site performance, with requests being rate-limited and the script designed in a such a way that the crawling operation would only need to be performed once.

[5] All figures in this article were generated in the open source statistical software R, using the igraph package.

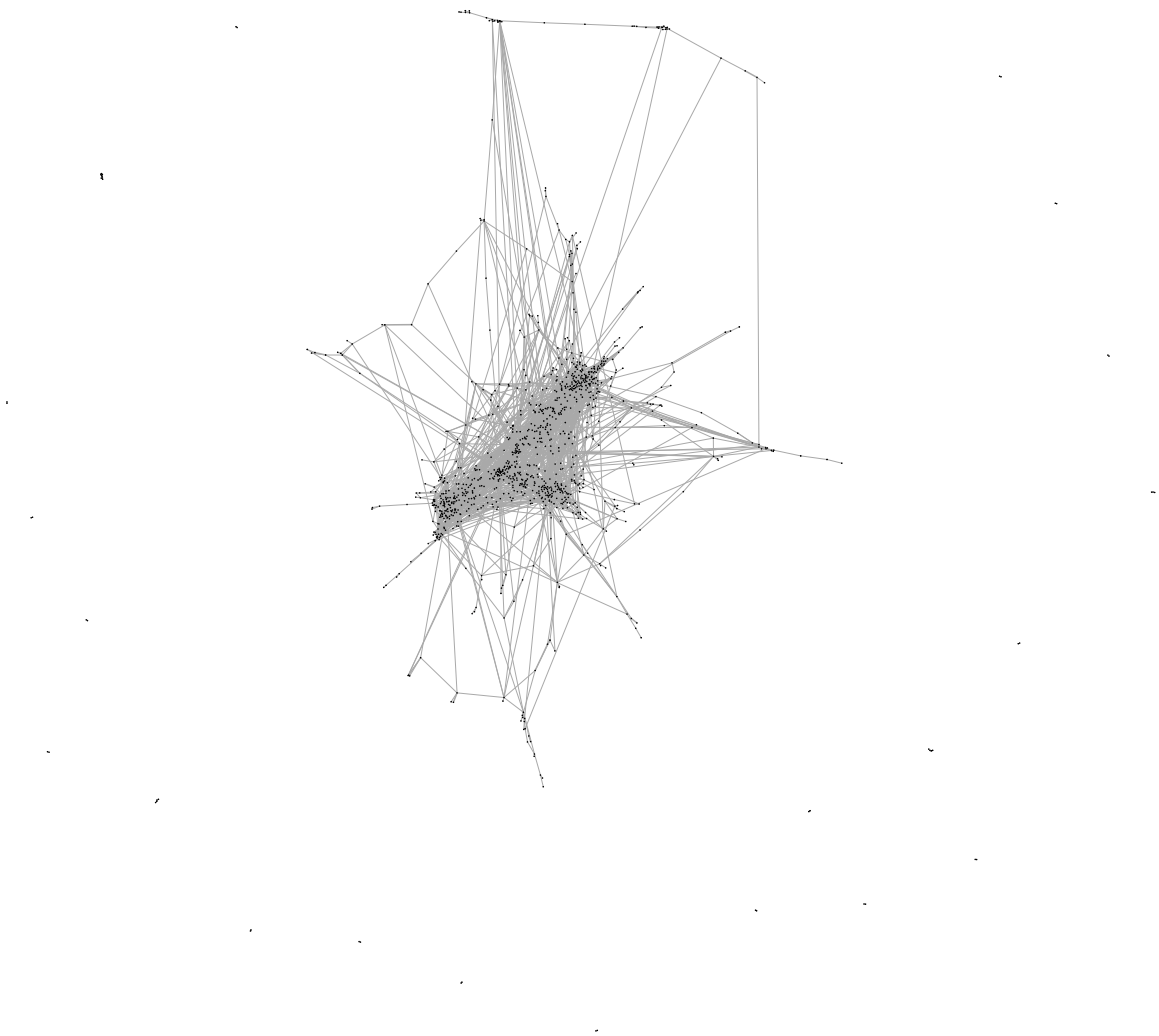

Figure 1: All links between entries (DrL layout)

than an evaluative term in this context—to the core of the *OCMP*. If literary groups are to be identified within that main stream, the graph as a whole must be cut down in a number of ways.

First, the small groups mentioned above can be removed: unattached to the main cluster, they are of no statistical use in the detection of groups within it. It is also possible to use the *OCMP*'s birthdate information to produce chronological subgraphs. Figure 2 shows poets born between 1910 and 1990, with links only drawn for edges with weights of 0.75 or higher, indicating that two poets were born relatively close together in time. At this level of restriction, the graph begins to be of some practical use.[6] A number of algorithms have been developed with the aim of detecting communities in graphs. One of these is *Walktrap*, which functions by taking short 'random walks' along a graph's edges.[7] The more closely connected a set of vertices (i.e. the higher the number of shared edges), the more likely it is that the walk will remain within it, identifying that set as a community. In Figure 3, *Walktrap* has used 3-step random walks to identify communities of at least 5 members in a modified version of the graph from Figure 2. Vertices with degrees of less than two—indicating poets connected to a group through only one other person—have been removed from the resulting subgraphs, but new one-degree vertices created by this process have been allowed to remain.

In order to test the effectiveness of *Walktrap* in identifying literary groups, it is necessary to develop some form of scoring mechanism.[8] A simple method is to compare the identified groups to 'actual' groups identified by the editors of the *OCMP* and given separate entries.[9] Any score (s) needs to reflect both positive accuracy—the number of members of a given group who are identified—and negative accuracy—how well the algorithm avoids misidentifying poets.[10] This can be done by adding the number of misidentified poets (m) to the total size of the human-defined group (h) and dividing the size of the algorithmically-defined group (a) by the result:

$$s = \frac{a}{m + h}$$

In the case of the Movement, the algorithm is somewhat conservative, correctly identifying 5 of the 9 poets listed in the *OCMP*'s entry—John Wain, D.J. Enright, Kingsley Amis, Elizabeth Jennings and Robert Conquest—and not misidentifying any. The score is therefore $\frac{5}{9+0} = 0.56$. With the Black Mountain poets, by contrast, it correctly identifies every member except Ed Dorn—an impressive 9 out of 10—but only at the cost of roping in a further 27 misidentified poets, ultimately scoring just $\frac{9}{10+27} = 0.24$. Turning to the New York School, the algorithm identifies 10 of the 14 poets listed as members or associates and misidentifies another 3, for a score of $\frac{10}{14+3} = 0.59$. Defining the notoriously nebulous Cambridge School, Nigel Wheale identifies 13 poets split into two 'generations'; *Walktrap* identifies 7 of these and misidentifies 5, scoring $\frac{7}{13+5} = 0.39$. With every group except the Movement, it is important to recognise that the 'misidentified' poets are frequently borderline or arguable members, such as Barry MacSweeney, or younger poets influenced by the main group, such as David Lehman.

Other communities identified by the algorithm do not conform clearly to literary groups listed in the *OCMP*, but they nevertheless represent recognisable clusters of poets: one of the largest communities, for example, is structured around a core of American poets typically described as 'confessional', including Sylvia Plath, Robert Lowell, Anne Sexton, John Berryman and W.D. Snodgrass. A much

[6] Zoom in to view labels clearly. This and the rest of the figures in this article are laid out using the Fruchterman–Reingold algorithm, which requires more computational power than DrL and is therefore unsuitable for large graphs such as that shown in Figure 1.

[7] Pascal Pons and Matthieu Latapy, 'Computing communities in large networks using random walks', *arXiv.org* (12 December 2005), <http://arxiv.org/abs/physics/0512106> [accessed 23 April 2016].

[8] Relying as it does on random walks, *Walktrap* is a non-deterministic algorithm—it may produce different results from run to run. In practice, variation is likely to be very small.

[9] The membership of the latter is calculated in the broadest possible way, as all poets mentioned in an entry.

[10] In the rare cases where the algorithm identifies more than one plausible candidate group, the group with the largest number of positive identifications is used, regardless of its overall size.

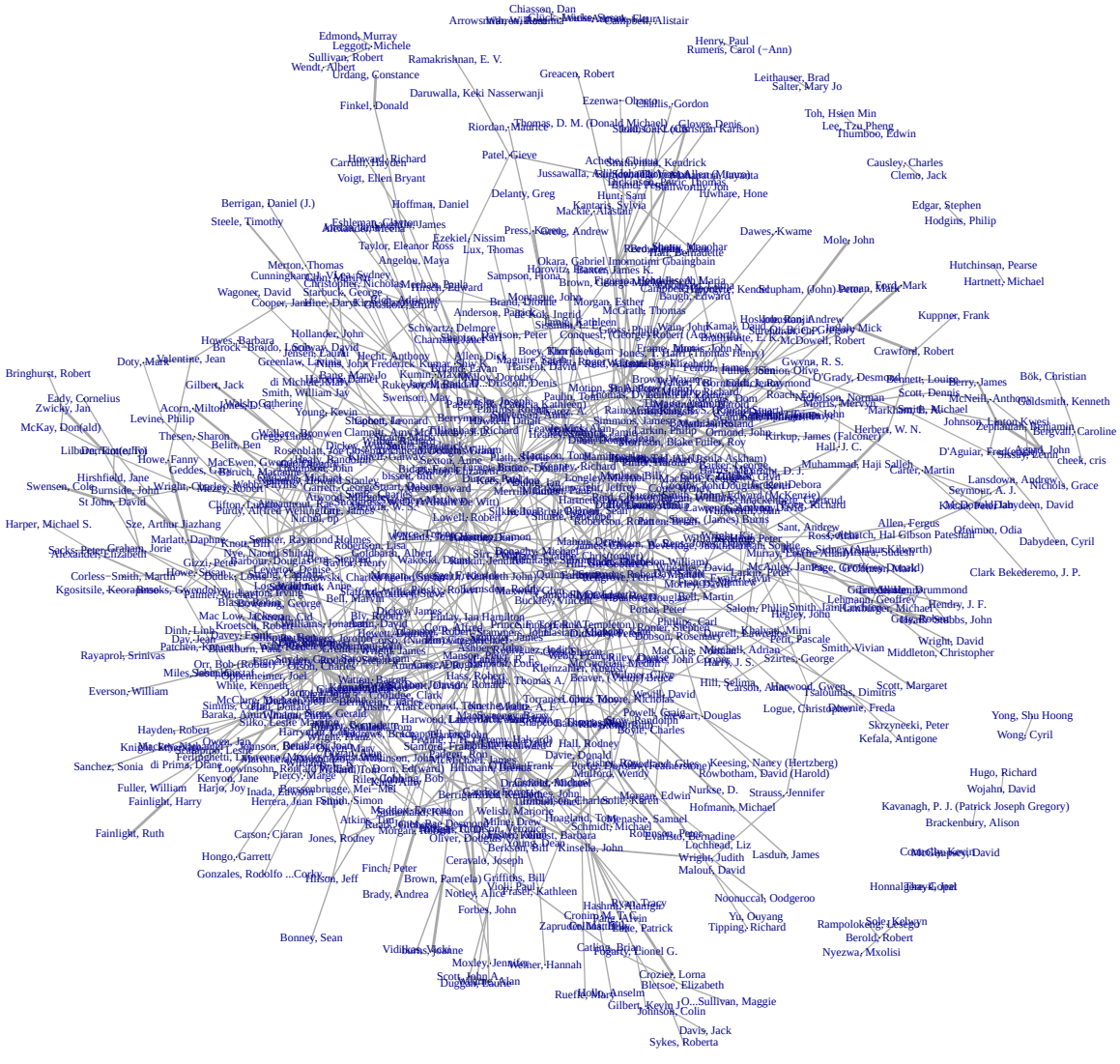

Figure 2: Poets born between 1910 and 1990 with connections of weight ≥ 0.75

Figure 3: Communities of at least 5 members identified by the *Walktrap* algorithm among edges of (weight ≥ 0.75

smaller cluster links together a number of quite diverse writers—Steve McCaffery, Michael Ondaatje, Margaret Atwood, bpNichol—united by their Canadian citizenship. Even when groupings do not appear to make sense from the perspective of current literary history, comparison can bring them into relief: while the position of, say, Tom Leonard and Bob Cobbing in the orbit of Language poets Charles Berstein and Bruce Andrews might at first seem odd, it might just as well testify to their eccentricity in relation to the mainstream of British poetry; it certainly *feels* more appropriate than positioning them in, say, the same group as Don Paterson, Roddy Lumsden, Michael Donaghy and Paul Farley. The key variable here is the weight cutoff below which edges are discarded, set at 0.75 in Figure 2. Modifying this variable radically changes the structure of the underlying graph, and therefore also the communities identified by *Walktrap*. Increasing it to 0.95, for example—meaning that poets must be almost exact contemporaries to be connected on the graph—makes for much smaller, and in some cases much more accurate, communities.

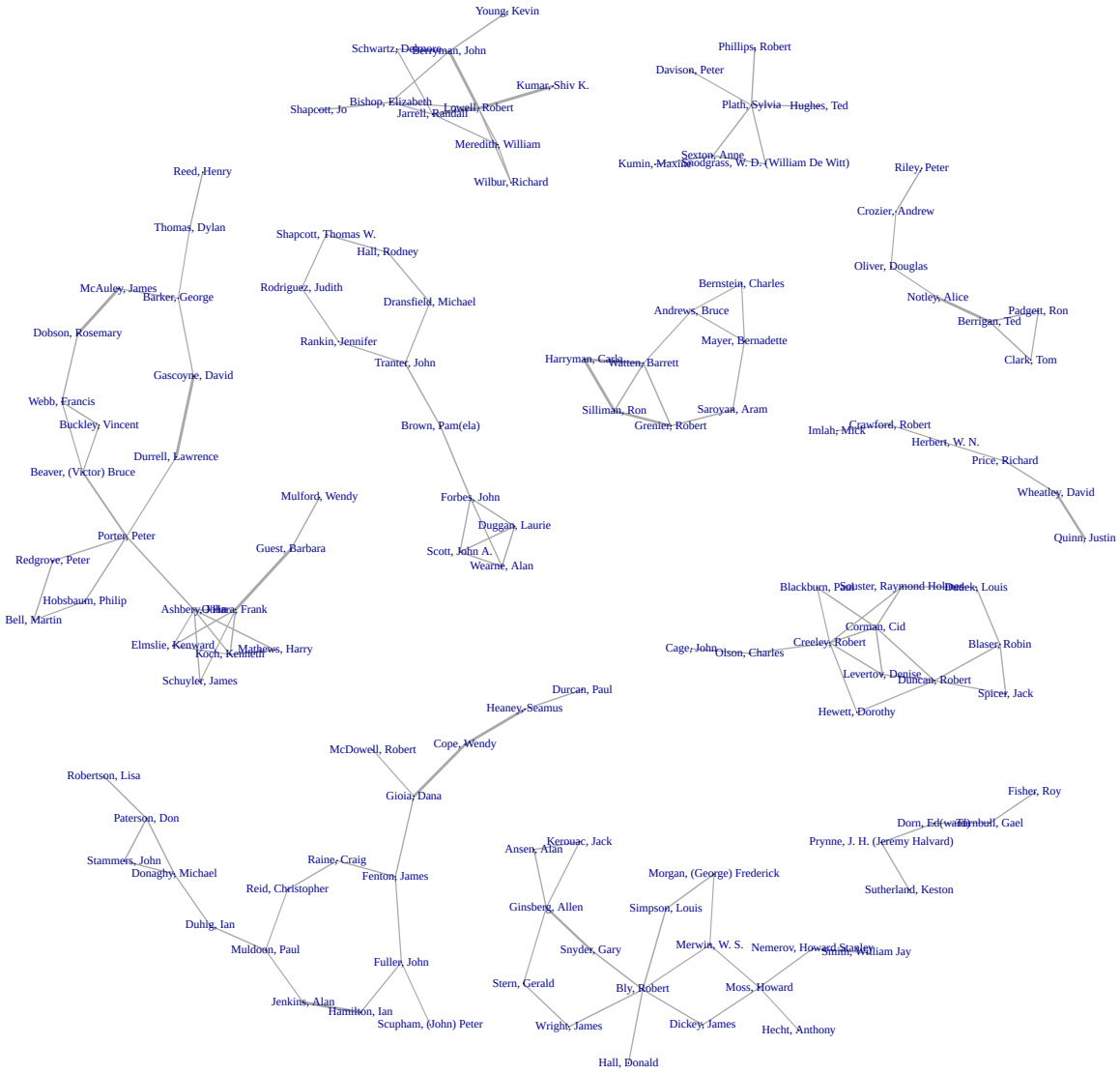

Figure 4: Communities of at least 5 members identified by the *Walktrap* algorithm among edges of weight ≥ 0.95

Table 1: Scores for *Walktrap* with various weight cutoffs

Weight cutoff	*Mov.*	*N.Y. Sch.*	*Cam. Sch.*	*Bl. Mntn.*	*Lang. Po.*	Average
0.60	0.45	0.15	0.26	0.11	0.06	0.21
0.65	0.56	0.59	0.28	0.24	0.05	0.34
0.70	0.56	0.59	0.33	0.24	0.05	0.35
0.75	0.56	0.59	0.39	0.24	0.15	0.39
0.80	0.56	0.50	0.30	0.27	0.18	0.36
0.85	0.36	0.58	0.12	0.29	0.22	0.31
0.90	0.36	0.23	0.16	0.33	0.50	0.32
0.95	0.00	0.24	0.18	0.29	0.56	0.25

In Figure 4, for example, the Black Mountain school is cut down drastically, and groups like the Language poets emerge much more clearly: *Walktrap* identifies five of the core group of six listed in Jeremy Noel-Tod's entry on Language Poetry, missing only Lyn Hejinian, and misidentifies—though in this case the term is hardly applicable—Bernadette Mayer, Aram Saroyan and Robert Grenier, for a score of $\frac{5}{6+3} = 0.56$. In some cases, however, increasing the weight cutoff creates false positives: poets who are only tangentially related by the standards of literary history but who happen to be of a similar age are linked, creating the characteristic 'strings' seen in Figure 4. Building on the scoring method detailed above, the mechanism at work here can be summarised as a relationship between three factors: *accuracy* (whether poets are correctly placed), *comprehensiveness* (how many poets are correctly identified in each group) and *definition* (how effectively groups are separated from each other). A 'loose' graph with a low weight cutoff sacrifices both accuracy and definition for comprehensiveness; a medium weight cutoff offers the highest definition and reasonable comprehensiveness at the price of accuracy; a 'tight' graph with a high weight cutoff generates communities with low comprehensiveness and occasional problems with definition, but provides the highest accuracy.

In the case of literary groups, empirical measurement—and especially optimisation—of a community detection algorithm is only possible to a very limited extent. In the first place, there is no central register of schools and movements against which findings can be tested: the membership lists used above are the work of individual scholars and would no doubt be disputed by others. At the same time, structure varies *between* groups. The threads binding the New York School together—primarily shared residence in New York City, but also a connection to the art world and to a lesser extent education at Harvard—are not the same as those which apply to the Black Mountain poets, a group centred very specifically on Black Mountain College. As Table 1 illustrates, it is unlikely to be possible to optimise an algorithm to make it better at extracting all groups from a graph: what boosts the accuracy, comprehensiveness or definition of one group will likely decrease that of another.[11] The real use of the algorithm is in fact to sharpen these distinctions, thereby revealing new information about each group. To take one example, reading the scores next to the communities' visual representation suggests that the solidity of the Movement and the Language poets as groups is the result of their compact but decentred structures: each member is connected, at most, to three others. This is notably different from, say, the Cambridge School or the Black Mountain poets, where J.H. Prynne and Charles Olson function as central hubs connected to large numbers of other poets. Yet the fact that the Movement and the Language poets do not show up reliably under the same weight cutoff conditions testifies to a difference in age ranges, with the latter being much closer contemporaries. Another disjunction can be hypothesised based on groups' respective influence: the more a literary group inspires followers, the harder it will be to identify algorithmically, being more closely connected

[11] Weight cutoffs below 0.6 produced uniformly similar results and were excluded from the table for clarity.

to other non-group points on the graph. The ideal literary group from an algorithmic perspective would be one which was completely self-contained, as with the detached groups discussed above.

For literary historians, the most interesting results from these techniques can be gained through a trial-and-error approach, modifying variables like the weight cutoff and the overall chronological range and paying attention to the resulting transformations in the communities detected. It is of course also possible to select individual poets or human-defined groups from the graph and to trace only their connections. One potential avenue for future research would be to identify an individual group such as the New York School and to test the efficacy of a number of different algorithms—igraph implements at least six, with a number of user-created implementations also available online—in picking it out of a larger graph. What this research ultimately depends on, however, is the availability of such graphs in the first place. Large-scale web crawling such as that described above exists in a legal grey area and is wasteful in terms of time and bandwidth, especially when replicated. The source code for this paper is available freely online; it would be helpful—not to mention in publishers' own interests—to apply the same principle to the link networks for reference volumes such as *The Oxford Companion to Modern Poetry*.[12] There is, after all, no significant copyright risk in revealing metadata, but its potential uses in humanities research are numerous.

[12]Full source code and data, excluding the crawler script, are available at https://github.com/ltrgoddard/lit-groups/.