International Journal of Arts & Sciences,
CD-ROM. ISSN: 1944-6934 :: 5(6):523–538 (2012)

WAY TO GO: A GRAPH MINING BASED APPROACH TO RECORD LINKAGE FOR SPARSE DATA

Alexander von Lünen

University of Portsmouth, UK.

Record linkage is an important field for Digital History and Historical Demography, among others. To link data from different recordsets existing algorithms almost always take at least two discriminating data into account, such as a name and date of birth, to link records of an individual. If there is only sparse data available for a person, i.e. only a name, it is usually dismissed as impossible to link by the record linkage literature. This paper introduces an algorithm to link sparse data sets on individuals when, by and large, only nominal data is available.

Keywords: Record linkage, Data mining, Historic data, Graph theory.

1. Introduction

Record linkage usually takes at least two discriminating descriptive items into account to link up different record sets, such as name and date of birth (Winchester 1992, 151). Data sets like census records normally provide for these quite readily and linking to these from other sets involves checking for name variants etc. to make sure a match is really that – a match. For some data sets the endeavor to link records appears an impossible task, especially when misspellings are frequent and little to no "non-nominative variables" (i.e. data other than names; Bouchard 1986, 15) is available to distinguish individuals.

This paper will outline the author's attempts to link such a sparse data set. With an improved fuzzy string matching method, a graph mining approach is taken to link record sets which carry numerous misspellings and few clues to identify individuals. The data set being used are the annual reports of the Steam Engine Makers' Society (SEM), a nineteenth century British trade union. The reports were digitized in the 1990s by a team at the University of London,[1] but some of the data in the reports escaped all linkage efforts, which led to the approach described in this paper.

The project of the author is designed to develop techniques to store, document and analyze social networks over space and time, and historic trade union records were chosen to be used as a

[1] The SEM database was originally developed by a team led by Dr Humphrey Southall, then with the Department of Geography at Queen Mary and Westfield College, University of London, UK. The team first digitized the annual reports between 1836 and 1876, and then tried to link individual members across years. See http://www.geog.port.ac.uk/lifeline/sem_db/sem_db_home.html for details, last accessed 11 Jan 2012.

testbed for the project. For that the digitized records of the SEM were downloaded from the UK Data Archive.[2] As it turned out, however, records that concerned the mobility of workers were not linked correctly and thus the most important feature for the author's research project was not usable.

Before the structure and nature of the data is explored, it would be a good idea to have a quick look at the SEM and its history to see the value of the data extracted from the annual report.

1.1 The Steam Engine Makers' Society

In 1824, after the *Combinations Act* from 1799/80 had been repealed that forbade the formation of trade unions in the UK, the *Steam Engine Makers' Society* was founded in Liverpool. In it, various trades concerned with the creation, erection and maintenance of steam engines combined to cope with a thriving market that required a mobile workforce. The main trades gathered in the union were fitters and turners, but also included millwrights and pattern makers, and later smiths, draughtsmen and "makers of tools generally used in the manufacture of steam engines" (Scotson 1865, 9). It was open to artisans from the age of 20 who had undergone an apprenticeship lasting between five and seven years.[3]

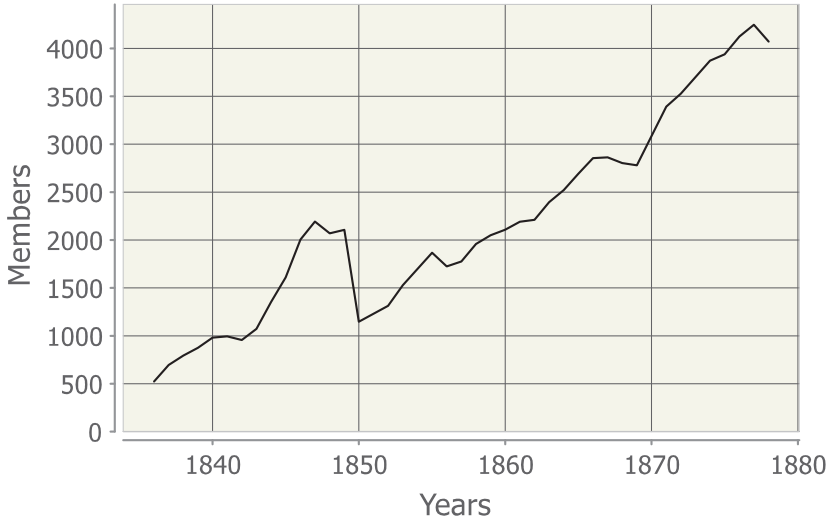


Figure 1. Membership figures for the SEM, compiled from the annual reports.

Around this time engineering trade unions were highly localized and rather small, which was due to the shift in economic structures. Where formerly the trade of millwrights covered most of engineering, industrialization demanded more specialized classes of workmen. Not only was steam engine technology geographically limited to the industrial hubs of Britain in the early 1800s, but smaller local trade unions were also considered to handle labour disputes in those hubs better (Marsh and Ryan 1984, 4). This was one of the reasons the SEM resisted the almagamation movement of 1851. By this time, several smaller unions had come to the

[2] http://www.esds.ac.uk/findingData/snDescription.asp?sn=3677&key=steam+engine+makers, last accessed 11 Jan 2012.

[3] Cf. SEM rulebooks 1827, 1846, 1865, 1879.

conclusion that the wider level of industrialization that had occurred demanded a greater movement to represent workers' interests better, and thus they formed the Amalgamated Society of Engineers (ASE) in 1851, uniting seven sectional engineering unions (Marsh and Ryan 1984, 4–6). The SEM at first undertook steps to be part of the almagamation of 1851, but suddenly withdrew from it, following disputes within the SEM. The motives for that decision are not entirely clear, but one contributing factor to the refusal was that branches feared that they would be outvoted and lose power if they were to become part of the bigger ASE (Marsh and Ryan 1984, 61). The SEM was highly specialized, and the research by this author suggests that steam engine ships were an important part of the SEM's business, a rather marginal market at the time over which they certainly would not want to lose their dominance by sharing it with other engineers.[4]

Consequently the SEM remained marginal in size during the nineteenth century, never exceeding 6,000 members (see figure 1). It grew more rapidly in the years between 1900 and 1914, after helping to found the Federation of Engineering and Shipbuilding Trades in 1891. Eventually, after a ballot of its members voted for further almagamation, it joined the ASE together with other unions in 1922 to form the Amalgamated Engineering Union (Marsh and Ryan 1984, 61).

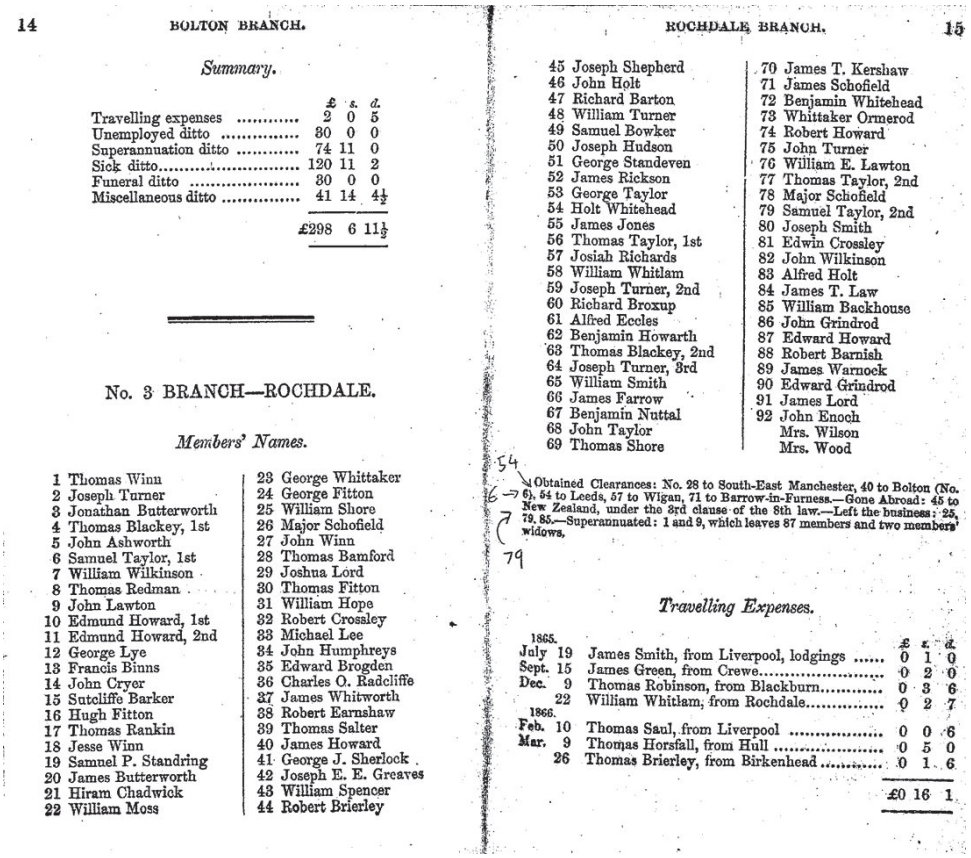
1.2 The Sources, i.e. the "Raw" Data

The only records and documents which have survived from the SEM are, by and large, the annual reports printed by the society, which were compiled from the monthly reports. Material for both were requested from each single branch by the central branch and were compiled into the annual report of the whole society. Each branch reported its members and expenditures etc., including payments of travel money, sick pay and superannuation. From these figures the financial and occupational situation of the SEM can be easily inferred.

There was no central register of members in place, let alone a central membership number.[5] Each branch kept records of its members and listed them in order of their seniority, i.e. the longer a member was with the particular branch, the higher up in the list would he be. If members with the same name happened to be member of the same branch, they were usually differentiated with "Jr" and "Sr" if they were related, or "1st" and "2nd" if they were not. Apart from some occasional inconsistencies this scheme worked quite well for a single branch. Difficulties arose where members changed branch, because it quite often was not clear which one had changed, as the next report would drop the name suffix from the remaining member.

[4] A full account of the history of the Steam Engine Makers' Society is beyond the scope of this paper, for more see for example: Hobsbawm (1951); or Southall (1991).

[5] At least no such register book was available in the early years of the society, thus rendering an analysis over time somewhat random.



14 BOLTON BRANCH.

Summary.

	£	s.	d.
Travelling expenses	2	0	5
Unemployed ditto	30	0	0
Superannuation ditto	74	11	0
Sick ditto	120	11	2
Funeral ditto	30	0	0
Miscellaneous ditto	41	14	4½
	£298	6	11½

No. 3 BRANCH—ROCHDALE.

Members' Names.

1 Thomas Winn
2 Joseph Turner
3 Jonathan Butterworth
4 Thomas Blackey, 1st
5 John Ashworth
6 Samuel Taylor, 1st
7 William Wilkinson
8 Thomas Redman
9 John Lawton
10 Edmund Howard, 1st
11 Edmund Howard, 2nd
12 George Lye
13 Francis Binns
14 John Cryer
15 Sutcliffe Barker
16 Hugh Fitton
17 Thomas Rankin
18 Jesse Winn
19 Samuel P. Standring
20 James Butterworth
21 Hiram Chadwick
22 William Moss
23 George Whittaker
24 George Fitton
25 William Shore
26 Major Schofield
27 John Winn
28 Thomas Bamford
29 Joshua Lord
30 Thomas Fitton
31 William Hope
32 Robert Crossley
33 Michael Lee
34 John Humphreys
35 Edward Brogden
36 Charles O. Radcliffe
37 James Whitworth
38 Robert Earnshaw
39 Thomas Salter
40 James Howard
41 George J. Sherlock
42 Joseph E. E. Greaves
43 William Spencer
44 Robert Brierley

ROCHDALE BRANCH. 15

45 Joseph Shepherd
46 John Holt
47 Richard Barton
48 William Turner
49 Samuel Bowker
50 Joseph Hudson
51 George Standeven
52 James Rickson
53 George Taylor
54 Holt Whitehead
55 James Jones
56 Thomas Taylor, 1st
57 Josiah Richards
58 William Whitlam
59 Joseph Turner, 2nd
60 Richard Broxup
61 Alfred Eccles
62 Benjamin Howarth
63 Thomas Blackey, 2nd
64 Joseph Turner, 3rd
65 William Smith
66 James Farrow
67 Benjamin Nuttal
68 John Taylor
69 Thomas Shore
70 James T. Kershaw
71 James Schofield
72 Benjamin Whitehead
73 Whittaker Ormerod
74 Robert Howard
75 John Turner
76 William E. Lawton
77 Thomas Taylor, 2nd
78 Major Schofield
79 Samuel Taylor, 2nd
80 Joseph Smith
81 Edwin Crossley
82 John Wilkinson
83 Alfred Holt
84 James T. Law
85 William Backhouse
86 John Grindrod
87 Edward Howard
88 Robert Barnish
89 James Warnock
90 Edward Grindrod
91 James Lord
92 John Enoch
Mrs. Wilson
Mrs. Wood

54
6 → 6), 54 to Leeds, 57 to Wigan, 71 to Barrow-in-Furness.—Gone Abroad: 45 to
7 New Zealand, under the 3rd clause of the 8th law.—Left the business: 25,
79 73, 85.—Superannuated: 1 and 9, which leaves 87 members and two members'
widows.

Travelling Expenses.

		£	s.	d.
1865.				
July 19	James Smith, from Liverpool, lodgings	0	1	0
Sept. 15	James Green, from Crewe	0	2	0
Dec. 9	Thomas Robinson, from Blackburn	0	3	6
22	William Whitlam, from Rochdale	0	2	7
1866.				
Feb. 10	Thomas Saul, from Liverpool	0	0	6
Mar. 9	Thomas Horsfall, from Hull	0	5	0
26	Thomas Brierley, from Birkenhead	0	1	6
		£0	16	1

Figure 2. Members and travel expenses for the Rochdale branch from the SEM Annual Report 1865/66.

The two females at the end of the members list are widows of former SEM members, and received a benefit payment from the branch.

The ranking in the branch helped to mitigate this shortcoming considerably. As mentioned, the rank in the branches' membership list indicated their seniority. It could therefore be concluded that a member would only climb up the ranks, or remain at the same rank, as he progressed his membership. While this pattern was sometimes broken by a branch (for no apparent reason), a member's rank would eventually help to identify the member in ambiguous cases, as outlined below.

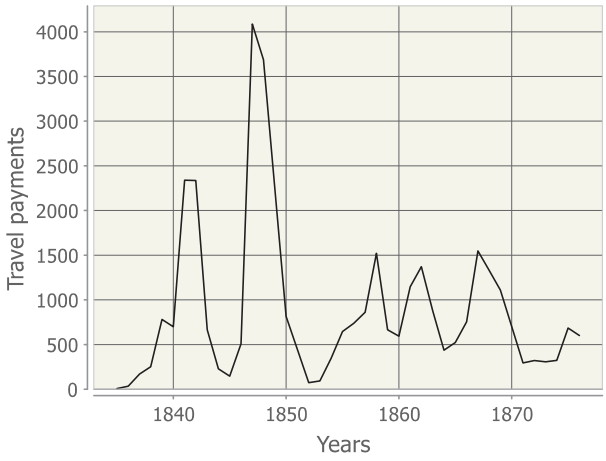


Figure 3: Travel figures for the SEM, compiled from the annual reports.

While getting the members from each annual report linked was a challenge, the linkage of the traveling information almost proved insolvable.[6] As mentioned above, every branch listed the expenditures for their fiscal year, travel payments among them. These travel payments were usually not made for members of that branch, but for those SEM members traveling from other branches to find work. The reports then give information such as "John Smith, Bolton, 1s 1d" for one branch (e.g. Liverpool) to indicate that 1s 1d had been paid to John Smith for traveling from Bolton to Liverpool. While John Smith not a member of the Liverpool branch (or at least not a John Smith from that branch, but from another one), the name "Bolton" only means that this was the last leg on his tramping. The rulebooks of the SEM clearly stated that travel payments were only made for the trip from the nearest branch, i.e. a member could not claim travel payment for the whole journey from, say, Southampton to Liverpool from the latter. Instead he had to go to the nearest branch to ask for available jobs and get the travel payment for the recent leg. If no jobs were available at this branch or its vicinity, the member would travel to the next, and so on until he either found employment or had traversed all branches. If he then still hadn't found a job he would return to his home branch and be entitled to unemployment benefit. If a member found work other than through the branch he had to report this to the nearest branch or pay a fine; also, he was not allowed to reject an offer of a job made by the branch, other than for cogent reasons.

The travel information thus recorded did not provide any more information beyond the name of the member and the last leg of his journey. This situation was worsened by the organizational and cultural practice of trade union business in those days. The annual reports usually list the locations and addresses of the branches in the appendixes. A quick survey of these addresses reveal that the great majority of branch offices were located in pubs, and meetings usually held on Friday nights, i.e. the branch secretary would sit in a pub on a Friday night and wait for members to arrive and report to him to get their travel money. It is therefore likely that the secretary was not in the most sober state when members approached him, and the many spelling errors in the names recorded bear witness to that. In short: while the information as such (only a name, but no "home" branch) in the traveling payments hinders a linkage quite severely already, the many errors in spelling render an attempt to link these payments to individual members almost futile.

Another issue was that members frequently changed branches or were given clearance to join another branch when they found work in another town, for example. Sometimes the reports name the branch the member changed to, but quite often no such information was given, just the fact that a member was given clearance or joined another branch. To check which branch that member went to, all other branches had to be checked for a new member (i.e. low in the ranks) with the same name in the same or the following year, depending on whether the change took place close to the editorial deadline of the report.

1.2.1 Traveling Data

In the table with traveling information, the situation was even worse than in the members table. The first analysis was to check the linking for all entries in the traveling table with code 'T' (for 'to', i.e. when a member was deliberately sent *to* a place by his home branch). The linking should be comparably straight-forward, since it is obvious in these cases that the member was from the same branch that made the payment. Disregarding a few suffix ambiguities, i.e. where a

[6] Most members had already been linked correctly by the original team in London. Where mobility was involved, however, such as members changing their branch or members traveling, their algorithms failed.

branch had, say, a Jr and a Sr of the same name, but did not specify which one was sent out, the matching should be rather easy.

Yet, when the member IDs (given by the 'old' linkage software) in the traveling table were checked against the ones in the members table it turned out that only 18.28% were linked, i.e. had the identical identifier in both the travelers and the members table. For code 'F' (for 'from', i.e. a member traveled *from* another branch to the one making the payment) this was almost 0%, i.e. almost none of the code 'F' rows in the travelling table were linked to the member's table correctly.

According to the remarks in the annual reports of the SEM the travel information section in the reports could be categorized as in table 1, showing also the distribution of those categories over the rows in the data.

Table 1. Distribution and meaning of traveling codes in the SEM db, derived from the statement in the report.

Code	Meaning	Percentage
A	*Accommodation*, i.e. only a room for over-night stays was paid for	3.56%
T	*To*, i.e. a member was specifically sent to a situation and thus started the travel at his home branch	8.94%
F	*From*, i.e. a member was paid travel money for traveling from another branch/place to the branch that made the payment	84.43%
R	*Round-Trip*, i.e. the member received travel payment for traveling to a situation and back to the branch	0.77%
X	*Misc.*, i.e. cases that were not specified in greater detail, or no specification at all	3.56%

For the research project the *from* records were the most important ones, as the project wanted to track the mobility of the artisans. Records with code 'T' were usually rather short single trips, whereas code 'F' records showed longer journeys when collated. It is thus quite obvious that the correct linkage of code 'F' rows in the traveling table were vital to the project.

2 The "new" SEM database

While the goal of the author's project was not to create a "better" (i.e. more accurately linked) version of the SEM database (db), a "clean-up" of it was inevitable. After an initial assessment of the database, however, it became obvious that the linkage had to be improved, since the change of locations in particular was poorly linked, as outlined above.

2.1 Linkage Methodology

Given that there is little to no non-nominative data, it is obvious that string matching algorithms will play a major part in the linking of records within the SEM db. The nature of the data thus determines the linking strategy contemplated. Bouchard (1986) poignantly stated that there is a "great diversity of existing methods" due to varying "circumstance and goals of automatic record linkage" (Bouchard 1986, 9). Furthermore, he rightfully points out that it is impossible to link all records in a dataset (Bouchard 1986, 10). The strategy of the author to link the records in the SEM db was consequently driven by the goal to match as many records as can be feasible.

The linkage methodology follows the approach laid out by Wrigley and Schofield (1973), but with considerable improvements, as will be pointed out below. The main difference, again, is the availability (or lack thereof) of non-nominal data. Wrigley and Schofield (1973) deal with family reconstitution and therefore look at various records from parish registers. While these are

certainly difficult to link, they usually give more variables that can be used in the linkage process. They use graph visualizations (without calling them graphs) to illustrate the linkage problems and the algorithms to overcome them. More resemblance with graph theory is introduced by their "cluster formation" which corresponds to the construction of the travel graph below. Wrigley and Schofield discard records that have only nominal data, thereby leaving it out of their "cluster" (Wrigley and Schofield 1973, 79). The author's algorithm goes further than that, namely the merging of subgraphs and the exploitation of sparse non-nominal data.

Another inspiration was the paper by Skolnick (1973), in which he refers to "decision making techniques" that he attributed to the field of Artificial Intelligence (Skolnick 1973, 109). These decision trees are by now mainstream computer science and are part and parcel of data mining algorithms.[7] The criteria to identify a record-pair from the traveling records and the member records is effectively a decision tree, the only difference being that decision trees usually work with induction, whereas the author's approach allows for abductive reasoning through his weighting system, which does not ultimately classify a match in the records, but rather documents its plausibility. The decision tree will be demonstrated later. Firstly, it makes sense to shortly introduce and summarize the improved fuzzy string technique the author employs to find matches in the travel record.

2.2 Name Matching

Since string matching would be such a crucial point in the linkage of the SEM's records, the author surveyed existing fuzzy string algorithms to find out whether there is one better suited for this task than the Soundex method, which is still the most frequently used name matching technique.[8] The main reason for this is the fact that Soundex was specifically designed to match surnames in census reports (at least for names spoken in the English tongue), whereas most other algorithms are designed to match strings on a character-by-character basis. Although there are promising fuzzy string matching techniques available, the author found that the Soundex method, while inhibiting certain deficiencies, was the most suitable one if its shortcomings could be amended. One way to accomplish this is to use the algorithm by Levenshtein, often referred to as "edit distance". In short, Levenshtein's technique is a metric that measures how many edit operations, such as substitutions or deletions, are required to transform the string *A* into the string *B*. The number of operations involved to do this is the "edit distance" and the greater the number of operations the more different the two strings are. The author found that any value above three in the edit distance would make the two strings too different to represent a resemblance – i.e. the two strings being identical, but with spelling errors.
In the light of the survey of string matching algorithms, it was decided to use a variety of those algorithms (in this order):

1. Direct match of forename, surname and suffix;
2. Direct match of forename and surname; suffix not given;
3. Combine Soundex and Levenshtein; Levenshtein distance must be smaller than three, Soundex difference must be greater than three;

---

[7] Cf. Kantardzic (2003). A "tree" is a special case of a graph; namely "a tree is a connected graph that has no cycles". Ore (1990, 37) A popular example are family trees in genealogy.

[8] That survey is published in a different journal, which was not fixed by the time this paper went to press. For a general survey on fuzzy string matching techniques, see Cohen, Ravikumar and Fienberg (2003).

4. Levenshtein distance must be smaller than three, first character in both surnames must be identical;
5. Check visually/manually.

All of the above were applied accumulatively, i.e. if one step could not find a match for all records, the next step would be tried on those unmatched. For example, if direct matches (first with, then without suffix) could only be made for a limited number of records, the combined Soundex-Levenshtein approach would be tried on the remaining ones, and so on.

2.3 Travel Linking

Whereas the member's table could be re-linked with the help of auxiliary information provided in the reports, such as branch and seniority in that branch, the re-linking strategy in the traveling table had to take a different approach. This was first and foremost due to the fact that not only names were frequently misspelled, but also that member's could *not* be backtracked in their entirety. As outlined above, the travel regime of the artisans organized in the SEM would lead one to expect to follow the travel route of a particular member backwards in time and to eventually lead the algorithm to the home branch of that SEM member (where he was supposed to start his journey from, according to the SEM rulebook). It turned out, however, that this worked for very few of the travelers marked with code 'F'.

2.3.1 The Principal Idea

The SEM annual reports list the expenses for every branch, the travel payments among them. The records in each report specify the member's name, the place traveled from or to, and the amount paid to the member. The travel can thus be deduced from the branch making the payment and the place from/to which the member tramped. As mentioned above, members had to travel from the nearest branch. In principle this should allow one to backtrack the complete travel by listing the individual legs. The original project therefore employed a backtracking algorithm to start from the last leg in the journey and trace it back to its place of origin.[9]

In practice, however, the rule of traveling to the closest branch has rarely been adhered to. Furthermore, a quick analysis of the travel records revealed gaps in journeys in which either a leg of a journey was not recorded or the member had apparently decided to skip the nearest branch. A continuous journey that would be traceable back to the member's home branch is thus rather rare. In most cases, records of journeys were fragmentary, rendering a backtracking algorithm futile; furthermore, the already mentioned spelling errors hindered an efficient linkage of travel records with membership records. It is consequently not surprising that the original travel linking is very incomplete.

[9] A "backtracking" algorithm is an algorithm that attempts to find a solution by trying a variety of choices, usually modeled as a graph, and if no solution can be found, it *backtracks* to the last known partial solution and tries a different path from that node in the graph (cf. Black 2008). Backtracking algorithms are very popular and have been applied to a great variety of problems. The origin for this approach is unclear, but one of the first publications was done by Golomb and Baumert (1965); see Skiena (2008, 231ff) for a newer discussion of backtracking algorithms, among others.

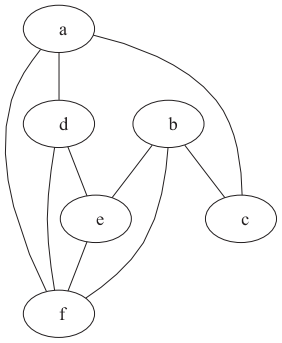


Figure 4. Example of a graph with $V = \{a,b,c,d,e,f\}$ and $E = \{(a,c),(a,d),(a,f),(b,c),(b,e),(b,f),(d,e),(d,f),(e,f)\}$.

Given this shortcoming of the sources, a different strategy of linkage was sought. First, an improved version of the backtracking solution was contemplated, but the results were rather disappointing and introduced more ambiguities than it solved. An algorithm based on the properties of graphs was then devised. As a matter of fact, one can find a similar approach in data mining literature, called "graph mining".[10] Usually used for a different domain of applications than record linkage, the general principle – especially when applied to Social Network Analysis – proved quite inspiring for this paper.

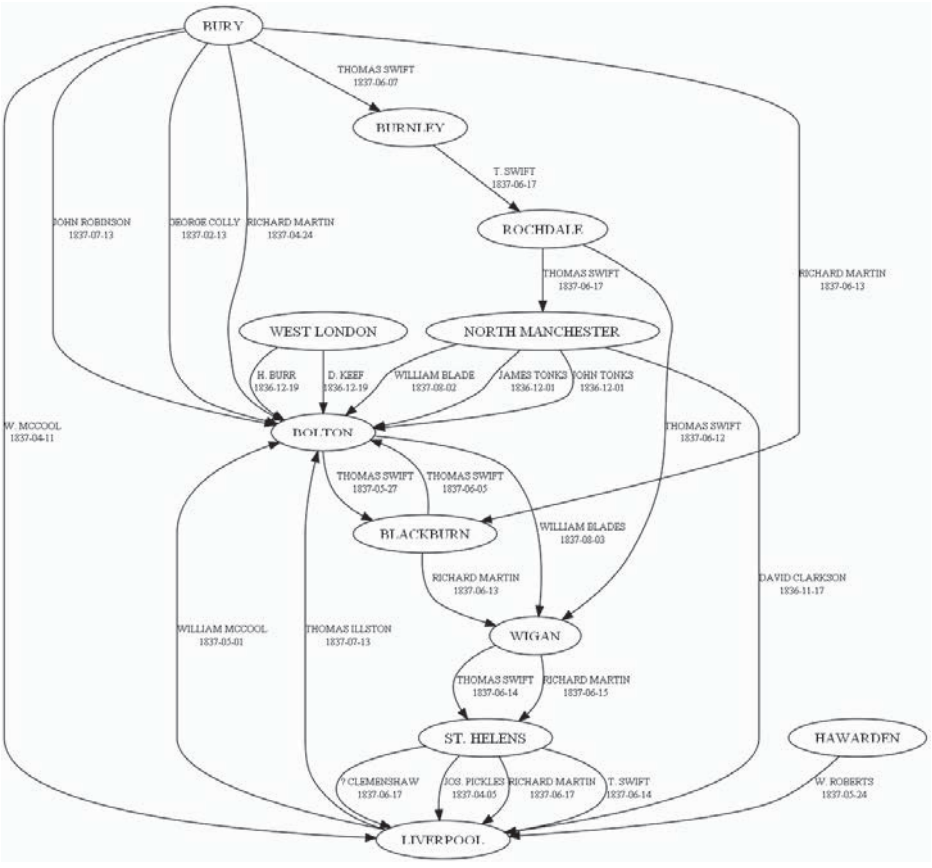


Figure 5. All travel records from 1836/37 in a graph.

[10] Cf. Han and Kamber (2006, 535ff); see also Washio and Motoda (2003), Katsaros (2009) for surveys of graph mining techniques.

2.3.2 Graph Theory to the Rescue

In this respect a brief introduction to graph theory would be in order. A graph G is an algebraic structure (not to be confused with the "graph" of a function) such that $G = (V,E)$, with V a set of vertices and E a set of edges. An "edge" is a connection of two vertices, i.e. for $e_i \in E$ and $v_l, v_k \in V$ we define $e_i = (v_l, v_k)$. The example in figure 4 shows an *undirected* graph, i.e. a graph in which the order of vertices that constitute an edge can be symmetric, i.e. (v_1, v_2) is the same as (v_2, v_1). For traveling movements this would not be the case, as the travel is from an origin to a goal (a return travel counts as two single travels). For this, graph theory offers *directed* graphs in which the symmetry of undirected graphs is not allowed. An example of a directed graph for traveling is shown in figure 6. Figures 5 and 6 also show the notion of *connected* graphs. The graph in figure 5 is a connected graph: for each vertex there exists a path to any other vertex, i.e. each vertex is connected to all other vertices, either directly or indirectly (i.e. via other vertices). Figure 6 is not a connected graph, as there is a portion that is unconnected to the rest of the graph. The graph in figure 6 consists of two *connected components*, i.e. two portions of the graph that are connected only locally.[11]

To link the traveling records to the membership information identical and similar names in the travel records were grouped and the journeys thus obtained modeled as a graph. If the graph turned out to be connected, checks were carried out to see whether the person in the travel payments could really be considered one and the same, i.e. checks were done to see if there was more than one person with that name (see below). If the graph was not connected, the same checks were applied to each subgraph (i.e. the unconnected portions of the graph), and tested to see whether the subgraphs concerned the same person.

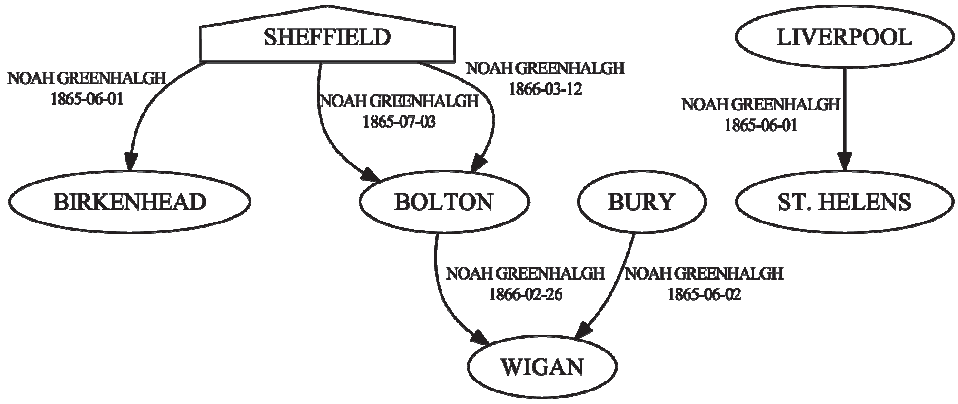


Figure 6. Traveling of Noah Greenhalgh in 1865/66 (house shape indicating his home branch), records producing an unconnected graph with two components.

2.3.3 Graph Mining

The author's actual research deals with social networks over space and time, more than record linkage or data mining. However, a sub-field of data mining – graph mining – offers some very inspiring pointers to a record linkage algorithm. Due to the growing popularity of Social

[11] Graphs are one of the most fundamental concepts in computer science and discrete mathematics. Therefore, a plethora of literature exists. For this paper, the author used Ore (1990) and Bondy and Murty (2008).

Network Analysis (SNA) in recent years a new field within graph mining has emerged, the so-called "link mining".

Han and Kamber discuss "several new tasks" that link mining brings about, compared to graph mining (Han and Kamber 2006, 560ff). Of these, the most applicable to the author's record linkage algorithm is that of "link-based object classification" (LBOC). Graph mining is chiefly about classifications of objects, i.e. the vertices in a graph, and clusters are usually detected by common attributes of those objects. In LBOC, however, the relationship between objects is explored and exploited in the mining process to gain more information. Classification of objects in a LBOC approach is thus not based on the attributes of an object alone, but also on the link and the attributes of the object being linked to (Han and Kamber 2006, 561).

There is one decisive difference, however, between the LBOC usually met in link mining/SNA and the author's approach: where LBOC and SNA is chiefly about exploring links between different objects (i.e. persons), in this LBOC application to record linkage, the objects on both side of a link are (supposedly) the same person; or rather: the LBOC approach by the author is used to establish that both objects are semantically the same person. The attributes (name, branches reported to on the journey, etc.) is hereby looked at in order to assess the plausibility that several travel records by a person with the same name (or a highly similar one) are all about the same person, or if they are about different persons with the same name.

The travel record, as outlined above, consist of the two endpoints of the journey (usually SEM branches, although this was not always the case) and the name of the person traveling between these two endpoints. To link the name of the traveler to a name in the membership records, one would have to check whether the person was a member of any of the branches involved on the journey. The LBOC was therefore taking the name (i.e. the edge/link between two branches) and the attribute data of the two vertices (i.e. the branches and their membership records) into account to establish a link between records.

2.3.4 Linking Algorithm

For this paper, therefore, the complete network is constructed from all traveling records in an annual report between branches by SEM members (i.e. the vertices in the graph are the SEM branches, and the (labeled) edges are the traveling records, i.e. the name of the member doing the travel and the date of the travel; see figure 5). In the beginning a name from the list of travelers in a report was picked at random. For every such name, the graph of all travelers in the report is traversed, trying to establish whether the person in question is a member of any of the two endpoints of a travel (i.e. origin and destination branches, or simply the two vertices forming the edge that is being inspected). If none of the endpoints is the member's home branch, this edge is regarded as a leg of a longer journey of a member. This boils down the problem of linking the traveling records with the membership records to a (fuzzy) string matching task with some logical constraints aiding the task. See figure 7 for details.

This procedure of traversing the graph is repeated for every name until all legs of the member's journey have been put into separate subgraphs. If a subgraph is unconnected, the following is applied to each connected component. At the end of the traversal, the home branch should have been found in one of the endpoints in the travel records.[12] If no such home branch

[12] Incidentally, it could happen that more than one home branch is found, as the member might have changed the branch. This was not so rare, since the member's journey might have resulted in a longer-term job elsewhere, and he

could be inferred from these endpoints, it is checked whether the member was listed at any branch of this year at all, and whether it could be decided with some authority that that branch member is the traveling member.

The algorithm rests on the assumption that connected portions of the traveling graph represent the same member, rather than two members with the same name. While not foolproof, it is highly unlikely that two or more members with the same name have traveled along the same branches at roughly the same time.[13] Given the extremely vague information on the travelers, some degree of uncertainty is inevitable, and a "weighting system" should document how high the level of uncertainty is, by adding up factors such as the score from the fuzzy string matching function or the ambiguity involved (e.g. whether the legs of the journey could be assigned to a member with utmost confidence); such a weighting helps the historian to assess and discuss how reliable the information is.

2.3.5 Weighting Systems

To decide whether two records produce a match, the record linkage literature usually refers to "weighting systems" that introduce a notion of "additivity of evidence […] combined with the notion that some combinations of items are less/more likely than others" (Winchester 1992, 156). The process of deciding whether a record in the traveling table should be included in a member's graph, i.e. the grouping of names to construct the traveling graph, involves such weighting, as discussed.

As Winchester (1992) outlines, there are essentially two weighting systems: an additive system, and one based on a likelihood ratio (Winchester 1992, 156ff).[14] Both Winchester (1992) and Skolnik (1973) favor the "simple addition of weights" (Winchester 1992, 156) over the likelihood ratio method. In the latter, a sample of record pairs is evaluated visually to assess the probability of correct and incorrect linkage, and this sample is then used to derive the probability of correctness for all records. Not only is this quite selective and may lead to a bias in the record linkage, but this also introduces an extra – and somewhat unnecessary – step into the process compared to the additive weighting method.[15]

Winchester (1970, 120ff) details different additive weighting systems. For the SEM db the author uses the (normalized) scores from the string matching algorithms. As outlined above, the record linkage in the SEM db relies almost exclusively on nominal data, since this is the only data given. It therefore makes sense to use the string matching score and add straight-forward scores to it, according to additional information such as rank in the branch.The scores as used by the author are listed in table 2.

---

consequently changed to the branch closer to his new workplace. Such changes are documented in the report, however, so these cases can be catered for by the algorithm quite easily.

[13] Unfortunately, the dates in the travel records did not prove to be very helpful. They record the date the branch made the payment to the member, rather than when the member traveled. Together with said poor record keeping skills by many branches' secretaries, this would create rather unreliable dates.

[14] In this he is confirming Skolnik (1973, 93ff) in his evaluation.

[15] Or "preferential scoring system", as Skolnik (1973, 94) calls it.

Table 2. Weights used in the linking algorithm.

Weight	Value
Fuzzy string matching	[0,1]
Branch membership unambiguous?	0.1 if yes
Rank in branch consistent?	0.1 if yes
Records in the same report?	0.1 if in the same report, 0.05 if in the following one, 0 if not
Travel records form a connected graph?	0.1 if yes, reduced by 0.01 for every unconnected component

It should be noted, however, that the author used the weights mostly for documentation purposes. The decision to link record pairs was decided in stages, as out lined above, and not upon the score given by the weighting system.

2.3.6 Merging Unconnected Components

The criteria to link the unconnected components of a travel graph is surely the most difficult part of the algorithm. In trivial cases, such as a member with a unique name or just one home branch across the unconnected components, the decision to link the unconnected components to that one member is rather uncontroversial. If these conditions are not met, however, linking the components can become a bit arbitrary, as it is usually not clear which parts of a journey refer to which member. After a short analysis – and during the linking process – it turned out that at maximum only two to three unconnected components were encountered per name and report. Almost all of these cases could be solved by checking for the home branch and the member's name, and the remaining ambiguous ones were so few that they could be ignored. Generally speaking, it was observed that journeys in geographically and temporally limited scopes were almost always done by the same person. This corresponded with the analysis that most members did not travel far from their home branch (i.e. remained fairly local), and did very few journeys on average.

3. Conclusions

Due to time constraints, the record linkage in the SEM db had to be stopped to carry out the actual research project. The author's algorithm could have been further improved by considering further logical constraints and hints. For example, an analysis revealed that quite often travel payments were made to persons who were not listed as members of any branch in the same report, but who appeared as a member in the next report. With further checking, such as uniqueness of the name, linking those records across reports would have been possible.

From the first run of the developed algorithm, however, satisfactory numbers of linked records had been achieved, so the linking was stopped. Of the traveling records with code 'T', 70.2% were linked (compared to the 18.28% from the 'old' SEM db) at this stage, and 52.4% of those records with code 'F' (compared to the almost 0% from the 'old' database). This provided quite a sufficient sample for the historic network analysis that the project was contemplating.

Most, if not all, literature on record linkage assumes the existence of two attributes for each record: a name and some auxiliary data such as date of birth, i.e. a person is uniquely identified by these parameters. In the dataset used by the author only one of these parameters existed: the name of the person. Linking such sparse datasets is extremely difficult, but not entirely impossible, with the help of combining graph mining techniques with fuzzy string matching and logical constraints. The presented approach can still be significantly improved by analyzing the

SEM db further to detect patterns to introduce more logical constraints to gain clues for potential matches.

The results, on the other hand, achieved by the algorithm developed by the author were very satisfying and linked more recordsets than most historians would regard as necessary for a sufficient sample.

Acknowledgements

This research was funded by the Leverhulme Trust under research project grant A20090049: "Data Models for Actor-Networks in Historical GIS: Workers and localities in C19". Warmest thanks go to Humphrey Southall, now Reader in Geography at the University of Portsmouth, UK, for his help and support in this research project. The original dataset of the Steam Engine Maker's Society was deposited with the UK Data Archive, and the newly linked records will be uploaded to it in due course.

For every report (1836-1876) **do**

1) Create a list with distinct names (forename/surname/suffix) from all travel records in this report;
2) Create a graph of travel records (1);
3) From the list of names (1), pick a name with no abbreviations in it and having the latest date;
4) Traverse the graph from (2) and get all connected nodes; delete them from the graph (2) and put them in a subgraph;
5) Check whether there are any more edges for that name (1); if yes, put them into a a second component of the same subgraph (4);
6) **If** the graph (4/5) is connected, **then**
   a) check the root vertex, i.e. the vertex that has only outgoing edges; this should be the home branch of the member (check by fuzzy string match);
   b) **if** there is no root vertex or the root vertex is not the member's home branch **then**
      i. check all other vertices in the graph, one of them should be the home branch; when more than one home branch is encountered, check whether member changed branch;
      ii. **if** none of the vertices is the home branch of the member, **then** check all branches for a member with that name (in this and the next report); **if** more than one candidate is encountered and no disambiguation is possible, **then** discard this member, i.e. leave travel record unlinked;
7) **Else** (i.e. the graph has unconnected components) **do** (6) for all unconnected components with the following alterations:
   a) **if** only one home branch could be established (by the procedure outlined in (6)), **then** assume the components concern the same person;
   b) **if** more than one home branch is encountered (and the member has not changed the branch), **then** mark the components for visual inspection; it is algorithmically difficult to decide whether this might indicate that different members are involved, and which home branch belonged to whom.

Figure 7. Algorithm for linking the travel records to the member records in the SEM db.

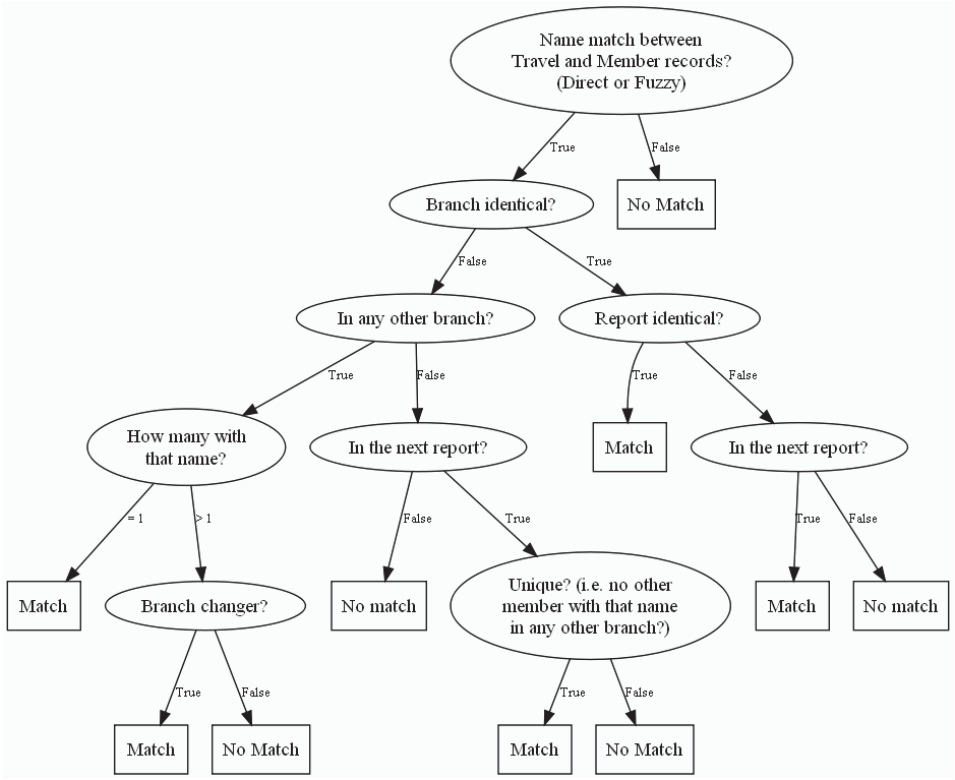


Figure 8. Decision tree for the algorithm.

References

1. Black, Paul E. 2008. Dictionary of Algorithms and Data Structures [online]. U.S. National Institute of Standards and Technology. Last accessed: 18 May 2011. http://xlinux.nist.gov/dads//HTML/backtrack.html.
2. Bondy, J.A. and U.S.R. Murty. 2008. *Graph Theory*. Graduate Texts in Mathematics. Berlin & Heidelberg: Springer.
3. Bouchard, Gérard. 1986. The Processing of Ambigous Links in Computerized Family Reconstruction. *Historical Methods* **19**(1): 9–19.
4. Cohen, William W., Pradeep Ravikumar, and Stephen E. Fienberg. 2003. A Comparison of String Distance Metrics for Name-Matching Tasks. In *Proceedings of IJCAI-03 Workshop on Information Integration,* edited by Craig Knoblock and Subbarao Kambhampati. Acapulco, Mexico, pp. 73–78.
5. Golomb, Solomon W. and Leonard D. Baumert. 1965. Backtrack Programming. *Journal of the ACM* **12**(4): 516–524.
6. Han, Jiawei and Micheline Kamber. 2006. *Data Mining: Concepts and Techniques*. 2nd ed. Burlington: Elsevier.
7. Hobsbawm, Eric. 1951. The Tramping Artisan. *The Economic History Review (New Series)* **3**(3): 299–320.
8. Kantardzic, Mehmed. 2003. *Data Mining: Concepts, Models, Methods, and Algorithms*. New York: John Wiley & Sons.
9. Katsaros, Dimitrios. 2009. Tree and Graph Mining. In *Encyclopedia of Data Warehousing and Mining*, edited by John Wang. 2nd ed. London: Information Science Reference, pp. 1990–1996.

10. Marsh, Arthur and Victoria Ryan. 1984. *Historical directory of trade unions*. Vol. 2: including unions in Engineering, Shipbuilding and Minor Metal Trades; Coal Mining and Iron and Steel; Agriculture, Fishing and Chemicals. Aldershot: Gower.
11. Ore, O. 1990. *Graphs and their Uses*. 2nd ed. Washington, D.C.: The Mathematical Association of America.
12. Scotson, Joseph (ed). 1865. *Laws of the Steam Engine Makers' Society*. Manchester: James Collins & Co.
13. Skiena, Steven S. 2008. *The Algorithm Design Manual*. 2nd ed. London: Springer.
14. Skolnick, Mark. 1973. The resolution of ambiguities in record linkage. In Wrigley (1973), pp. 102–127.
15. Southall, Humphrey. 1991. Mobility, the artisan community and popular politics in early nineteenth-century England. In: *Urbanising Britain: Essays on class and community in the nineteenth century,* edited by Gerry Kearns and Charles W. J. Withers. Cambridge Studies in Historical Geography. Cambridge: Cambridge University Press.
16. Washio, Takashi and Hiroshi Motoda. 2003. State of the art of graph-based data mining. *SIGKDD Explor. Newsl.* **5**(1): 59–68.
17. Winchester, Ian. 1970. The Linkage of Historical Records by Man and Computer: Techniques and Problems. *Journal of Interdisciplinary History* **1**(1): 107–124.
18. Winchester, Ian. 1992. What Every Historian Needs to Know About Record Linkage for the Microcomputer Era. *Historical Methods* **25**(4): 149–165.
19. Wrigley, E. A. (ed). 1973. *Identifying People in the Past*. London: Edward Arnold.
20. Wrigley, E. A. and R. S. Schofield. 1973. Nominal record linkage by computer and the logic of family reconstitution. In Wrigley (1973), pp. 64–101.