Using Metadata to find Paul Revere

2013-06-09 · Sociology · IT · Politics · Data · R

London, 1772.

I have been asked by my superiors to give a brief demonstration of the surprising effectiveness of even the simplest techniques of the new-fangled *Social Networke Analysis* in the pursuit of those who would seek to undermine the liberty enjoyed by His Majesty's subjects. This is in connection with the discussion of the role of "metadata" in certain recent events and the assurances of various respectable parties that the government was merely "sifting through this so-called metadata" and that the "information acquired does not include the content of any communications". I will show how we can use this "metadata" to find key persons involved in terrorist groups operating within the Colonies at the present time. I shall also endeavour to show how these methods work in what might be called a *relational* manner.

The analysis in this report is based on information gathered by our field agent Mr David Hackett Fischer and published in an Appendix to his *lengthy report to the government*. As you may be aware, Mr Fischer is an expert and respected field Agent with a broad and deep knowledge of the colonies. I, on the other hand, have made my way from Ireland with just a little quantitative training—I placed several hundred rungs below the Senior Wrangler during my time at Cambridge—and I am presently employed as a junior analytical scribe at ye olde National Security Administration. Sorry, I mean the Royal Security Administration. And I should emphasize again that I know nothing of current affairs in the colonies. However, our current Eighteenth Century beta of PRISM has been used to collect and analyze information on more than two hundred and sixty persons (of varying degrees of suspicion) belonging variously to seven different organizations in the Boston area.

Rest assured that we only collected *metadata* on these people, and no actual conversations were recorded or meetings transcribed. All I know is whether someone was a member of an organization or not. Surely this is but a small encroachment on the freedom of the Crown's subjects. I have been asked, on the basis of this poor information, to present some names for our field agents in the Colonies to work with. It seems an unlikely task.

If you want to follow along yourself, there is a secret repository containing the data and the appropriate commands for your portable analytical engine.

Here is what the data look like.

Code

```
                        StAndrewsLodge LoyalNine NorthCaucus LongRoomClub TeaParty Bosto
Adams.John                           0         0           1            1         0
Adams.Samuel                         0         0           1            1         0
```

```
4   Allen.Dr                0       0       1       0       0
5   Appleton.Nathaniel      0       0       1       0       0
6   Ash.Gilbert             1       0       0       0       0
7   Austin.Benjamin         0       0       0       0       0
8   Austin.Samuel           0       0       0       0       0
9   Avery.John              0       1       0       0       0
10  Baldwin.Cyrus           0       0       0       0       0
11  Ballard.John            0       0       1       0       0
12
13
```

The organizations are listed in the columns, and the names in the rows. As you can see, membership is represented by a "1". So this Samuel Adams person (whoever he is), belongs to the North Caucus, the Long Room Club, the Boston Committee, and the London Enemies List. I must say, these organizational names sound rather belligerent.

What can we tell from these meagre metadata? This table is large and cumbersome. I am a pretty low-level operative at ye olde RSA, so I have to keep it simple. My superiors, I am quite sure, have far more sophisticated analytical techniques at their disposal. I will simply start at the very beginning and follow a technique laid out in a beautiful paper by my brilliant former colleague, Mr Ron Breiger, called "The Duality of Persons and Groups." He wrote it as a graduate student at Harvard, some thirty five years ago. (Harvard, you may recall, is what passes for a university in the Colonies. No matter.) The paper describes what we now think of as a basic way to represent information about links between people and some other kind of thing, like attendance at various events, or membership in various groups. The foundational papers in this new science of social networke analysis, in fact, are almost all about what you can tell about people and their social lives based on metadata only, without much reference to the actual content of what they say.

Mr Breiger's insight was that our table of 254 rows and seven columns is an *adjacency matrix*, and that a bit of matrix multiplication can bring out information that is in the table but perhaps hard to see. Take this adjacency matrix of people and groups and transpose it—that is, flip it over on its side, so that the rows are now the columns and *vice versa*. Now we have two tables, or matrices, a 254x7 one showing "People by Groups" and the other a 7x254 one showing "Groups by People". Call the first one the adjacency matrix $\mathbf{A}$ and the second one its transpose, $\mathbf{A}^{\mathrm{T}}$. Now, as you will recall there are rules for multiplying matrices together. If you multiply out $\mathbf{A}(\mathbf{A}^{\mathrm{T}})$, you will get a big matrix with 254 rows and 254 columns. That is, it will be a 254x254 "Person by Person" matrix, where both the rows and columns are people (in the same order) and the cells show the number of organizations any particular pair of people both belonged to. Is that not marvelous? I have always thought this operation is somewhat akin to magick, especially as it involves moving one hand down and the other one across in a manner not wholly removed from an incantation.

I cannot show you the whole Person by Person matrix, because I would have to kill you. I jest, I jest! It is just because it is rather large. But here is a little snippet of it. At this point in the eighteenth century, a 254x254 matrix is what we call *Bigge Data*. I have an upcoming EDWARDx talk about it. You should come.

Code

```
1                           Adams.John Adams.Samuel Allen.Dr Appleton.Nathaniel
2   Adams.John              -           2           1                  1
3   Adams.Samuel             2          -           1                  2
4   Allen.Dr                 1           1          -                  1
```

```
5   Appleton.Nathaniel          1           2         1                  -
6   Ash.Gilbert                 0           0         0                  0
7   Austin.Benjamin             0           1         0                  0
8
```

You can see here that Mr Appleton and Mr John Adams were connected through both being a member of one group, while Mr John Adams and Mr Samuel Adams shared memberships in two of our seven groups. Mr Ash, meanwhile, was not connected through organization membership to any of the first four men on our list. The rest of the table stretches out in both directions.

Notice again, I beg you, what we did there. We did not start with a "social networke" as you might ordinarily think of it, where individuals are connected to other individuals. We started with a list of memberships in various organizations. But now suddenly we *do* have a social networke of individuals, where a tie is defined by co-membership in an organization. This is a powerful trick.

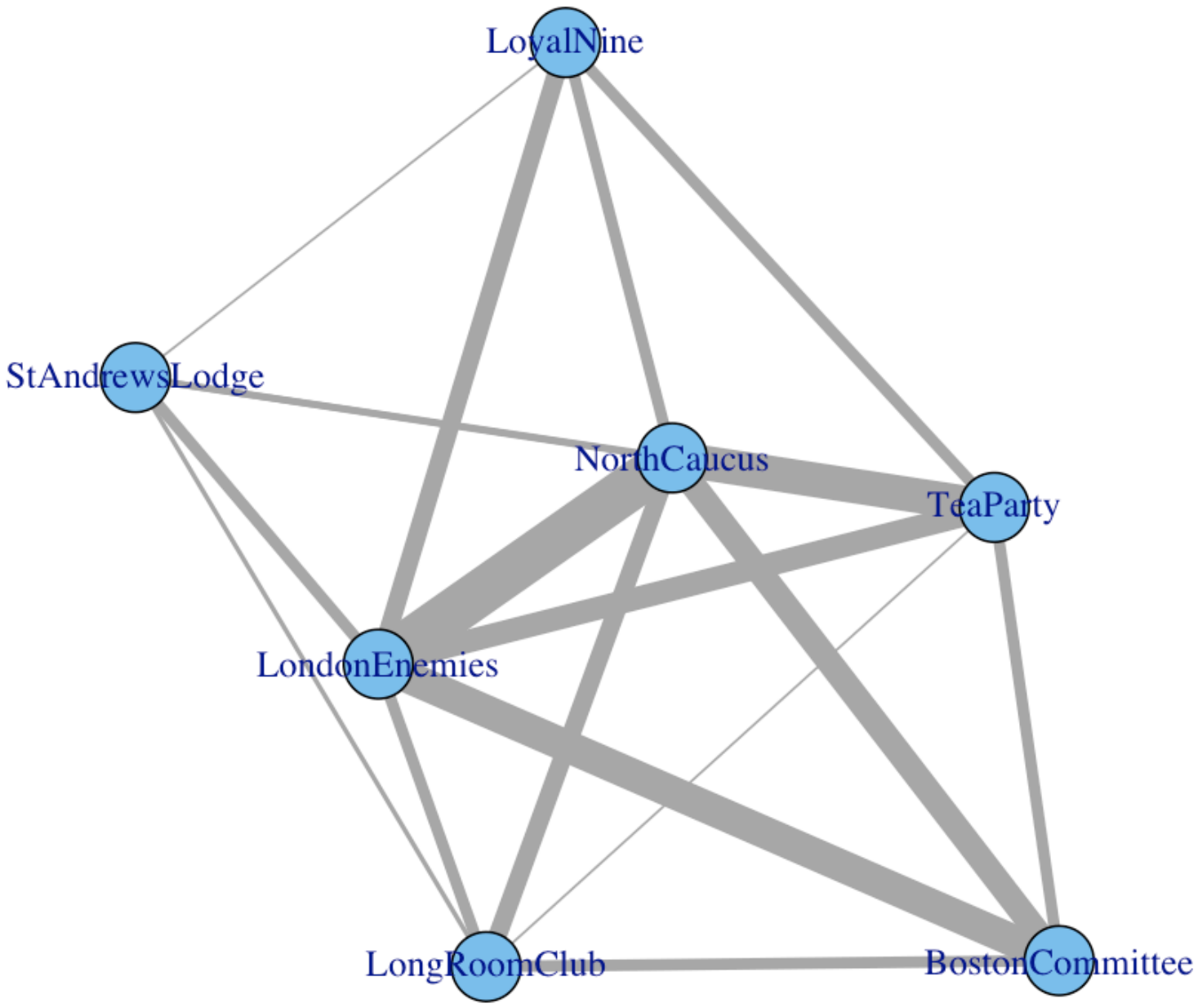
We are just getting started, however. A thing about multiplying matrices is that the order matters. It is not like multiplying two numbers. If instead of multiplying $\mathbf{A}(\mathbf{A}^{\mathrm{T}})$ we put the transposed matrix first, and do $\mathbf{A}^{\mathrm{T}}(\mathbf{A})$, then we get a different result. This time, the result is a 7x7 "Organization by Organization" matrix, where the numbers in the cells represent how many people each organization has in common. Here's what that looks like. Because it is small we can see the whole table.

Code

```
1                    StAndrewsLodge LoyalNine NorthCaucus LongRoomClub TeaParty BostonComm
2   StAndrewsLodge                -         1           3            2         3          0
3   LoyalNine                     1         -           5            0         5          0
4   NorthCaucus                   3         5           -            8        15         11
5   LongRoomClub                  2         0           8            -         1          5
6   TeaParty                      3         5          15            1         -          5
7   BostonCommittee               0         0          11            5         5          -
8   LondonEnemies                 5         8          20            5        10         14
9
```

Again, interesting! (I beg to venture.) Instead of seeing how (and which) people are linked by their shared membership in organizations, we see which organizations are linked through the people that belong to them both. People are linked through the groups they belong to. Groups are linked through the people they share. This is the "duality of persons and groups" in the title of Mr Breiger's article.

Rather than relying on tables, we can make a picture of the relationship between the groups, using the number of shared members as an index of the strength of the link between the seditious groups. Here's what that looks like.



The network of groups

And, of course, we can also do that for the links between the people, using our 254x254 “Person by Person” table. Here is what that looks like.

The network of individuals

What a nice picture! The analytical engine has arranged everyone neatly, picking out clusters of individuals and also showing both peripheral individuals and—more intriguingly—people who seem to bridge various groups in ways that might perhaps be relevant to national security. Look at that person right in the middle there. Zoom in if you wish. He seems to bridge several groups in an unusual (though perhaps not unique) way. His name is Paul Revere.

Once again, I remind you that I know nothing of Mr Revere, or his conversations, or his habits or beliefs, his writings (if he has any) or his personal life. All I know is this bit of metadata, based on membership in some organizations. And yet my analytical engine, on the basis of absolutely the most elementary of operations in Social Networke Analysis, seems to have picked him out of our 254 names as being of unusual interest. We do not have to stop here, with just a picture. Now that we have used our simple "Person by Event" table to generate a "Person by Person" matrix, we can do things like calculate centrality

scores, or figure out whether there are cliques, or investigate other patterns. For example, we could calculate a *betweenness centrality* measure for everyone in our matrix, which is roughly the number of "shortest paths" between any two people in our network that pass through the person of interest. It is a way of asking "If I have to get from person a to person z, how likely is it that the quickest way is through person x?" Here are the top betweenness scores for our list of suspected terrorists:

Code

```
round(btwn.person[ind][1:10],0)
     Revere.Paul     Urann.Thomas    Warren.Joseph      Peck.Samuel
            3839             2185             1817             1150
Barber.Nathaniel   Cooper.William     Hoffins.John      Bass.Henry
             931              931              931              852
    Chase.Thomas      Davis.Caleb
             852              852

```

Perhaps I should not say "terrorists" so rashly. But you can see how tempting it is. Anyway, look—there he is again, this Mr Revere! Very interesting. There are fancier ways to measure importance in a network besides this one. There is something called eigenvector centrality, which my friends in Natural Philosophy tell me is a bit of mathematics unlikely ever to have any practical application in the wider world. You can think of it as a measure of centrality weighted by one's connection to other central people. Here are our top scorers on that measure:

Code

```
> round(cent.eig$vector[ind][1:10],2)
Barber.Nathaniel     Hoffins.John   Cooper.William      Revere.Paul
            1.00             1.00             1.00             0.99
     Bass.Henry      Davis.Caleb    Chase.Thomas Greenleaf.William
            0.95             0.95             0.95             0.95
   Hopkins.Caleb   Proctor.Edward
            0.95             0.90

```

Here our Mr Revere appears to score highly alongside a few other persons of interest. And for one last demonstration, a calculation of Bonacich Power Centrality, another more sophisticated measure. Here the lower score indicates a more central location.

Code

```
> round(cent.bonpow[ind][1:10],2)
     Revere.Paul     Urann.Thomas    Warren.Joseph   Proctor.Edward
           -1.51            -1.44            -1.42            -1.40
Barber.Nathaniel     Hoffins.John   Cooper.William      Peck.Samuel
           -1.36            -1.36            -1.36            -1.33
    Davis.Caleb     Chase.Thomas
           -1.31            -1.31

```

And here again, Mr Revere—along with Messrs Urann, Proctor, and Barber—appears towards the top or our list.

So, there you have it. From a table of membership in different groups we have gotten a picture of a kind of social network between individuals, a sense of the degree of connection between organizations, and some

strong hints of who the key players are in this world. And all this—all of it!—from the merest sliver of metadata about a single modality of relationship between people. I do not wish to overstep the remit of my memorandum but I must ask you to imagine what might be possible if we were but able to collect information on very many more people, and also *synthesize* information from different *kinds* of ties between people! For the simple methods I have described are quite generalizable in these ways, and their capability only becomes more apparent as the size and scope of the information they are given increases. We would not need to know what was being whispered between individuals, only that they were connected in various ways. The analytical engine would do the rest! I daresay the shape of the real structure of social relations would emerge from our calculations gradually, first in outline only, but eventually with ever-increasing clarity and, at last, in beautiful detail—like a great, silent ship coming out of the gray New England fog.

I admit that, in addition to the possibilities for finding something interesting, there may also be the prospect of discovering suggestive but ultimately incorrect or misleading patterns. But I feel this problem would surely be greatly ameliorated by more and better metadata. At the present time, alas, the technology required to automatically collect the required information is beyond our capacity. But I say again, if a mere scribe such as I—one who knows nearly nothing—can use the very simplest of these methods to pick the name of a traitor like Paul Revere from those of two hundred and fifty four other men, using nothing but a list of memberships and a portable calculating engine, then just think what weapons we might wield in the defense of liberty one or two centuries from now.

Note: After I posted this, Michael Chwe emailed to tell me that Shin-Kap Han has published an [article analyzing Fischer's Revere data](http://www.sscnet.ucla.edu/polisci/faculty/chwe/ps269/han.pdf) in rather more detail. I first came across Fischer's data when I read *Paul Revere's Ride* some years ago. I transcribed it and worked on it a little (making the graphs shown here) when I was asked to give a presentation on the usefulness of Sociological methods to graduate students in Duke's History department. It's very nice to see Han's much fuller published analysis, as he's an SNA specialist, unlike me.

‹ Following up on Paul Revere

Updates to the Social Science Starter Kit ›