

Alla ricerca di Paul Revere con i metadati

May 21, 2017

Londra, 1772

I miei superiori mi hanno richiesto di fornire una breve dimostrazione della sorprendente efficacia delle più semplici tecniche di questa nuova e strana *Social Networke Analysis* nella caccia a coloro i quali vorrebbero tentare di minare la libertà di cui godono i sudditi di Sua Maestà. Questo è collegato alla discussione sul ruolo dei "metadati" per quanto riguarda certi avvenimenti recenti e le rassicurazioni di varie rispettabili parti in causache il governo stesse solamente "setacciando questi cosiddetti metadati" e che "le informazioni acquisite non includono il contenuto di alcuna comunicazione". Vi farò vedere come sia possibile utilizzare questi "metadati" per individuare le persone chiavi coinvolte in gruppi terroristici attualmente operanti nelle Colonie. Tenterò anche di mostrare come questi metodi funzionino in ciò che può essere chiamata una maniera *relazionale*.

Le analisi in questo rapporto si basano sulle informazioni raccolte dal nostro agente operativo Mr David Hackett Fischer e sono pubblicare in un'appendice al suo *esteso rapporto al Governo*. Come potreste sapere, il signor Fischer è un esperto e rispettato Agente operativo con una vasta e profonda conoscenza delle colonie. Io, d'altro canto, sono giunto dall'Irlanda con poco addestramento effettivo - mi sono qualificato in graduatoria con qualche centinaio di posti in meno rispetto al Senior Wrangler durante il mio periodo a Cambridge - e sono al momento impiegato come uno scriba analitico inferiore alla buon vecchia National Security Administration. Scusate, volevo dire la Royal Security Administration. E dovrei nuovamente sottolineare che non so nulla delle faccende correnti nelle colonie. Tuttavia, il nostro attuale beta 'XVIII Secolo' del PRISM è stato utilizzato per raccogliere e analizzare le

informazioni riguardanti più di 260 persone (di vari gradi di sospetto) appartenenti a sette diverse organizzazioni presenti nell'area di Boston.

Restiate assicurati che non è stata registrata alcuna conversazione né è stato riportato alcun incontro; sono stati raccolti solamente *metadati* su queste persone. Sicuramente questa è solo una piccola invadenza nei confronti della libertà dei sudditi della Corona. Mi è stato richiesto, sulle basi di queste scarse informazioni, di produrre qualche nome con cui possano lavorare i nostri agenti nelle Colonie. Sembrerebbe un incarico improbabile.

Se voleste seguire la procedura passo passo, eccovi la repository segreta contenente i dati e i comandi appositi per la vostra macchina analitica.

Ecco come si presentano i dati.

	St Andrew's Lodge	Loyal Nine	North Caucus	Long Room Club	Tea Party	Boston Committee	London Enemies List
Adams, John	0	0	1	1	0	0	0
Adams, Samuel	0	0	1	1	0	1	1
Allen, Dr	0	0	1	0	0	0	0
Appleton, Nathaniel	0	0	1	0	0	1	0
Ash, Gilbert	1	0	0	0	0	0	0
Austin, Benjamin	0	0	0	0	0	0	1
Austin, Benjamin	0	0	0	0	0	0	1
Avery, John	0	1	0	0	0	0	1
Baldwin, Cyrus	0	0	0	0	0	0	1
Ballard, John	0	0	1	0	0	0	0

Le organizzazioni sono riportate nelle colonne, e i nomi nelle righe. Come potete vedere, l'appartenenza a un gruppo è contrassegnata da un '1'. Dunque questo Samuel Adams (chiunque egli sia) appartiene al North Caucus, al Long Room Club, alla Boston Committee, e alla London Enemies List. Devo dire che queste organizzazioni hanno nomi alquanto belligeranti.

Cosa possiamo ricavare da questi miseri metadati? Questa tabella è grande e scomoda. Io sono un operativo di livello abbastanza basso alla buon vecchia RSA, quindi devo cercare di fare le cose semplici. Sono sicuro che i miei superiori abbiano a loro disposizione delle tecniche analitiche di gran lunga più sofisticate. Io mi limiterò semplicemente a cominciare dalle basi, seguendo una tecnica descritta in un bellissimo articolo da un mio precedente collega, Mr Ron Breiger, chiamato "La dualità delle persone e dei gruppi". Lo scrisse da laureato ad Harvard, circa trentacinque anni fa. (Harvard, se rammentate, è ciò che spacciano come università nelle colonie. È di poca importanza.) L'articolo descrive ciò che adesso consideriamo come una semplice maniera per rappresentare i legami fra persone e qualche altra sorta di cosa, come la partecipazione a vari eventi, o l'appartenenza a vari gruppi. Gli articoli alla base di questa nuova scienza di social network analysis sono fondati infatti quasi tutti su cosa si può dire riguardo le persone e la loro vita sociale basandosi unicamente sui metadati, senza troppi riferimenti al reale contenuto di ciò che esse dicono.

L'intuizione di Mr Breiger fu quella di capire che la nostra tabella di 254 righe e sette colonne è una *matrice delle adiacenze*, e che un po' di moltiplicazione matriciale può tirare fuori delle informazioni presenti nella tabella ma forse non troppo visibili. Prendete questa matrice delle adiacenze e trasponetela - ovvero giratela dall'altra parte in modo che le righe diventino le colonne e *viceversa*. Adesso abbiamo due tabelle, o matrici, una da 254x7 che mostra le "Persone per Gruppo", e una da 7x254 che mostra i "Gruppi per Persone". Siano la prima la matrice delle adiacenze \mathbf{A} e la seconda la sua trasposta, \mathbf{A}^T . Come vi potete ricordare vi sono delle regole per moltiplicare le matrici. Moltiplicando $\mathbf{A}\mathbf{A}^T$ otterrete una grossa matrice con 254 righe e 254 colonne. Sarà quindi una matrice 254x254 "Persona per Persona", dove sia le righe che le colonne sono persone (riportate nello stesso ordine) e le entrate mostrano il numero di organizzazioni a cui ogni coppia di persone appartiene. Non è meraviglioso? Ho sempre pensato che questa operazione fosse in qualche modo simile alla magia, soprattutto visto che si deve muovere una mano giù

e una lateralmente in maniera non troppo diversa da quella per un incantamento.

Non posso farvi vedere tutta la matrice Persona per Persona, perchè dovrei uccidervi. Scherzo, scherzo! È solo perchè è alquanto grande. Però eccovene una piccola porzione. A questo punto del Diciottesimo Secolo, una matrice 254x254 è ciò che viene chiamato *Bigge Data*. Terrò prossimamente su questo argomento un talk EDWARDx. Dovreste venire.

	Adams, John	Adams, Samuel	Allen, Dr	Appleton, Nahaniel	Ash, Gilbert	Austin, Benjamin
Adams, John	-	2	1	1	0	0
Adams, Samuel	2	-	1	2	0	1
Allen, Dr	1	1	-	1	0	
Appleton, Nathaniel	1	2	1	-	0	
Ash, Gilbert	0	0	0	0	-	0
Austin, Benjamin	0	1	0	0	0	-

Potete vedere qui che Mr Appleton e Mr John Adams sono collegati essendo entrambi membri di un gruppo, mentre Mr John Adams e Mr Samuel Adams fanno parte entrambi di due gruppi su sette. Mr Ash, invece, non è collegato a nessuno dei primi quattro nomi sulla lista tramite alcun gruppo. Il resto della tabella si estende in entrambe le direzioni.

Notate ancora, vi prego, cosa abbiamo fatto qui. Non abbiamo cominciato con un "social network" come lo potreste solitamente immaginare, attraverso il quale individui sono collegati ad altri individui. Abbiamo cominciato con una lista di appartenenza a varie organizzazioni. Ma adesso improvvisamente *abbiamo* un social network di individui, in cui il legame è definito attraverso l'appartenenza di due individui a una stessa organizzazione. Questo è un truccetto potente.

Stiamo solo cominciando tuttavia. Una delle regole della moltiplicazione matriciale è che l'ordine di moltiplicazione è importante. Non è come moltiplicare due numeri. Se invece di moltiplicare $\mathbf{A}A^T$ avessimo messo prima la matrice trasposta e avessimo fatto $A^T\mathbf{A}$, avremmo ottenuto un risultato differente. Questa volta, il risultato è una matrice 7x7 "Organizzazione per Organizzazione", nella quale le entrate rappresentano il numero di persone che le due organizzazioni hanno in comune. Ecco com'è fatta. Essendo piccola, possiamo vedere l'intera tabella.

	St Andrew's Lodge	Loyal Nine	North Caucus	Long Room Club	Tea Party	Boston Committee	London Enemies List
St Andrew's Lodge	-	1	3	2	3	0	5
Loyal Nine	1	-	5	0	5	0	8
North Caucus	3	5	-	8	15	11	20
Long Room Club	2	0	8	-	1	5	5
Tea Party	3	5	15	1	-	5	10
Boston Committee	0	0	11	5	5	-	14
London Enemies List	5	8	20	5	10	14	-

Di nuovo, interessante! (Oso dire.) Invece di vedere come gli individui sono collegati attraverso l'appartenenza a un'organizzazione, vediamo quali organizzazioni sono legate attraverso le persone che fanno parte di entrambe. Le persone sono collegate dai gruppi ai quali appartengono. I gruppi sono legati attraverso le persone che condividono. Questa è la "dualità delle persone e dei gruppi" a cui si riferisce Mr Breiger nel titolo del suo articolo.

Invece di avere a che fare con le tabelle, possiamo creare un'immagine delle relazioni fra i gruppi, usando il numero di individui in comune come indice della forza del legame fra i vari gruppi sediziosi. Ecco come si presenta.

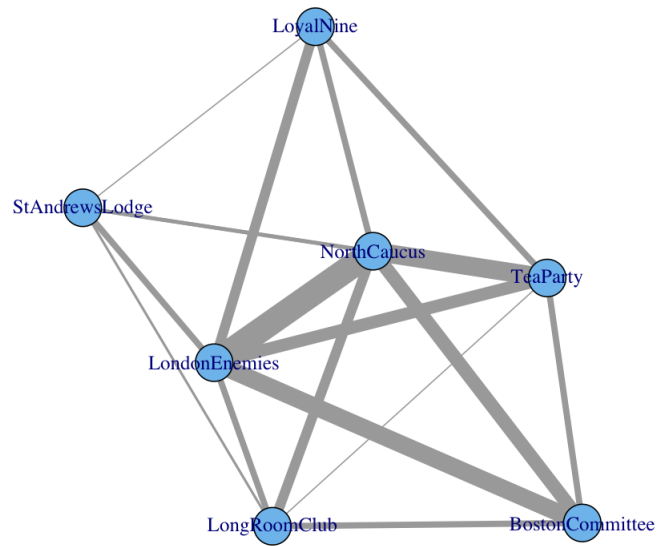


Figure 1: La rete dei gruppi

E ovviamente possiamo fare lo stesso per i legami fra le persone, usando la nostra tabella 254x254 "Persona per Persona". Ecco com'è fatta.

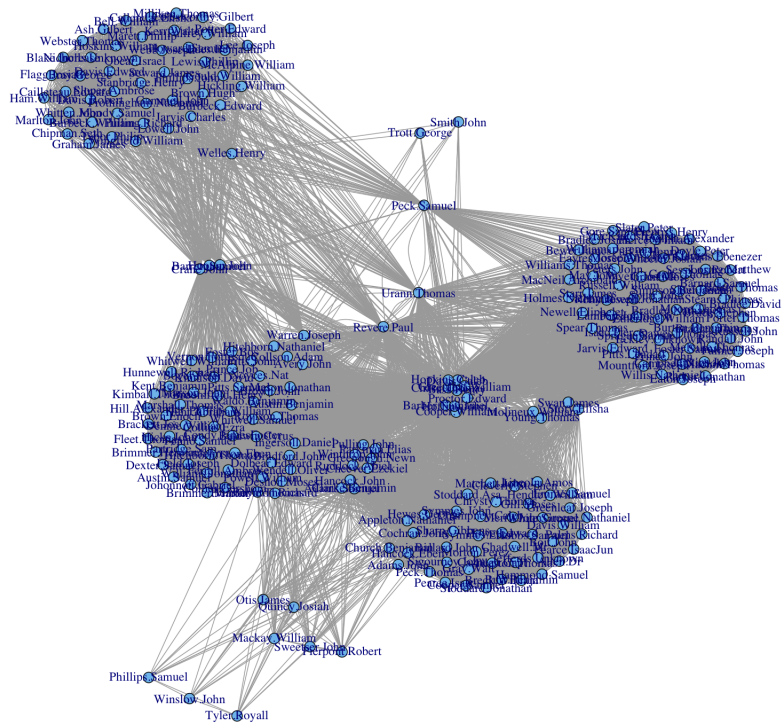


Figure 2: La rete degli individui

Che bella immagine! La macchina analitica ha disposto tutti in maniera ordinata, enfatizzando gli ammassi di persone e mostrando anche degli individui più periferici e - in maniera più interessante - quelli che sembrano agire da ponte fra più gruppi in modi che potrebbero essere rilevanti alla sicurezza nazionale. Guardate quella persona in mezzo. Ingrandite se volete. Costui sembra unire più gruppi in maniera insolita (seppure forse non unica). Il suo nome è Paul Revere.



Vi ricordo nuovamente che non so nulla del signor Paul Revere, delle sue conversazioni, delle sue abitudini o delle sue credenze, dei suoi scritti (se ne ha), o della sua vita privata. Tutto ciò che so proviene da questi scarsi metadati, basati sulla sua appartenenza ad alcune organizzazioni. Tuttavia la mia macchina analitica riesce, utilizzando le più basilari operazioni della Social Network Analysis, a individuarlo e a contrassegnarlo come una persona di inusuale interesse fra i nostri 254 nomi. Non dobbiamo fermarci qui, con solo una figura. Adesso che abbiamo utilizzato la nostra tabella "Persona per Organizzazione" per generare una matrice "Persona per Persona", possiamo fare cose come calcolare indici di centralità, oppure capire se si sono formate delle fazioni, o investigare altre tendenze. Per esempio, potremmo calcolare una misura di betweenness centrality per ogni persona nella nostra matrice. Questa è approssimativamente il numero di "cammini più brevi" da un individuo all'altro che passano attraverso la nostra persona di interesse. È un modo per chiedere "Se devo andare dalla persona a alla persona z, quanto è probabile che il cammino più veloce passi attraverso la persona x?". Ecco le misure più alte di betweenness per la nostra lista di sospettati terroristi:

Revere, Paul	3839
Urann, Thomas	2185
Warren, Joseph	1817
Peck, Samuel	1150
Barber, Nathaniel	931
Cooper, William	931
Hoffins, John	931
Bass, Henry	852
Chase, Thomas	852
Davis, Caleb	852

Forse non dovrei dire "terroristi" così incautamente. Ma potete vedere quanto possa essere allettante farlo. In ogni caso, guardate - eccolo di nuovo, questo Mr Revere! Molto interessante. Ci sono metodi più sofisticati per misurare l'importanza in un sistema. Vi è una cosa nota come centralità di autovettori, che i miei amici di Filosofia Naturale mi dicono appartenere ad una branca di matematica che probabilmente non avrà alcuna applicazione pratica. Potete pensarci come una misura di centralità pesata dal proprio collegamento ad altre persone centrali. Ecco le nostre misure più alte in questo caso:

Barber, Nathaniel	1.00
Hoffins, John	1.00
Cooper, William	1.00
Revere, Paul	0.99
Bass, Henry	0.95
Davis, Caleb	852
Chase, Thomas	852
Greenleaf, William	0.95
Hopkins, Caleb	0.95
Proctor, Edward	0.90

Qui il nostro Mr Revere sembra avere un punteggio alto insieme ad altre persone di interesse. E come ultima dimostrazione, un calcolo di Bonacich Power Centrality, un'altra misura sofisticata. Qui il punteggio più basso indica una posizione più centrale.

Revere, Paul	-1.51
Urann, Thomas	-1.44
Warren, Joseph	-1.42
Proctor, Edward	-1.40
Barber, Nathaniel	-1.36
Hoffins, John	-1.36
Cooper, William	-1.36
Peck, Samuel	-1.33
Davis, Caleb	-1.31

E qui nuovamente il signor Revere - insieme ai signori Urann, Proctor, e Barber - compare in cima alla nostra lista.

Così il gioco è fatto. Da una tabella di appartenenza a vari gruppi abbiamo ottenuto una rappresentazione di una sorta di social network fra individui, un'idea circa i collegamenti fra le varie organizzazioni, e qualche forte indizio sui giocatori chiave in questo ambiente. E tutto ciò - proprio tutto! - è stato ottenuto dalla più piccola porzione di metadati riguardanti una singola modalità di rapporti fra persone. Non vorrei oltrepassare i limiti del mio memorandum ma debbo chiedervi di immaginare cosa potrebbe essere possibile se solo potessimo raccogliere informazioni su molte più persone, e anche *sintetizzare* informazioni fra diversi *tipi* di collegamento fra persone! Perchè infatti i semplici metodi che ho delineato sono alquanto generalizzabili in queste maniere, e le loro capacità diventano solo più apparenti quando aumentano la quantità e il raggio d'azione delle informazioni che vengono fornite. Non dovremmo sapere ciò che confabulano due individui, ma solo che sono collegati in varie maniere. La macchina analitica farebbe il resto! Oserei dire che la vera struttura delle relazioni sociali emergerebbe gradualmente dai nostri calcoli, prima solo di profilo, ma in seguito con chiarezza sempre maggiore e, infine, in dettagli finissimi - come una grande, silenziosa nave che emerge dalla nebbia grigia del New England.

Ammetto che, oltre alle possibilità di trovare qualcosa di interessante, vi potrebbe anche essere il prospetto di scoprire andamenti suggestivi ma in realtà errati o fuorvianti. Tuttavia sono convinto che questo problema sarà sicuramente mitigato da metadati migliori e in maggiore quantità. Al momento, ahimè, la tecnologia necessaria per raccogliere automaticamente le informazioni necessarie è al di là delle nostre capacità. Ma ripeto, se un mero scriba come me - uno che sa quasi nulla - può utilizzare il più semplice di questi metodi per selezionare il nome di un traditore come Paul Revere tra quelli di altri 254 uomini, utilizzando solo una lista di appartenenze e una macchina analitica portatile, pensate a quali armi potremmo impugnare in difesa della libertà fra uno o due secoli.

1 Note su teoria dei grafi e analisi di reti sociali

1.1 Basi di teoria dei grafi

Un **grafo** G è una coppia di insiemi (V, A) detti rispettivamente **nodi** e **archi** del grafo. Gli archi sono il sottoinsieme di tutte le possibili coppie di nodi in V . Se queste coppie sono ordinate, allora si ha un **grafo orientato**; viceversa si ha un **grafo non orientato**. Graficamente, i nodi si rappresentano con i punti del piano e gli archi con segmenti (orientati se necessario) che collegano i nodi.

Ad esempio sia G il grafo costituito dall'insieme di nodi $V = \{a, b, c\}$. L'insieme degli archi sarà

$$A_o = \{(a, b); (a, c); (b, a); (b, c); (c, a); (c, b)\}$$

se il grafo è orientato e

$$A_{no} = \{(a, b); (b, c); (a, c)\}$$

se il grafo non è orientato.

Sia $(i, j) \in A$. Se il grafo G è orientato, si dice che i è il predecessore del nodo j (e viceversa j è il successore del nodo i). Se il grafo non è orientato, si dice che i due nodi sono adiacenti.

Si può inoltre costruire una **matrice delle adiacenze** avente come numero di righe il numero di nodi e come numero delle colonne il numero degli archi. Nella posizione (i, j) si ha un 1 se e solo se si ha un arco da i a j ; altrimenti si ha uno 0.

Un **cammino** è una sequenza di $m + 1$ nodi s in G tali che $\forall i = 1, \dots, m$ si ha che $(s_{i-1}, s_i) \in A$. m è la lunghezza del cammino. Un cammino è quindi una sequenza di archi di G a due a due.

Un nodo j viene detto accessibile dal nodo i se esiste un cammino da i a j .

Un cammino viene detto semplice se un arco viene percorso al più una volta, oppure elementare se ogni nodo viene percorso al più una volta.

Quando il primo e l'ultimo nodo di un cammino coincidono si ha un **ciclo**.

In un grafo orientato, un ciclo o un cammino possono essere a loro volta orientati se ogni arco viene percorso secondo il proprio orientamento.

1.2 Indici di centralità per analisi di reti sociali

- **Grado:** è la misura di archi per nodo, o di "popolarità" di un nodo. Se il grafo è orientato, per un nodo si possono definire l'indegree (archi in entrata) e l'outdegree (archi in uscita).
- **Closeness:** media delle distanze di un nodo da altri. Può essere quindi immaginata come la rapidità di propagazione dell'informazione da un nodo (può essere quindi utile per analizzare la diffusione di un virus all'interno di un grafo ad esempio).
- **Betweenness:** media del numero di volte in cui un nodo si trova in un cammino fra altri due nodi. È quindi una misura di quanto un nodo possa controllare l'informazione che scorre fra altri nodi, e permette di individuare il collegamento fra comunità che lavorano su fronti diversi - il nostro Paul Revere, ad esempio.

NB: il cammino (o la distanza) non si devono intendere come essere sempre minimi; anzi generalmente per una rete sociale non lo sono.

References

- [1] Istat. La Network analysis: uno strumento per lo studio delle reti. Rapporto annuale 2016.
- [2] E. Bozzo, D. Fasino, M. Franceschet. Slide da *Calcolo di indici di centralità di reti sociali*. Due Giorni di Algebra Lineare Numerica - Genova, 15-16 Febbraio 2012
- [3] Introduzione ai grafi
- [4] M.E.J. Newman. *Networks. An introduction*. Oxford University Press, New York, 2010.

2 Trasposizione

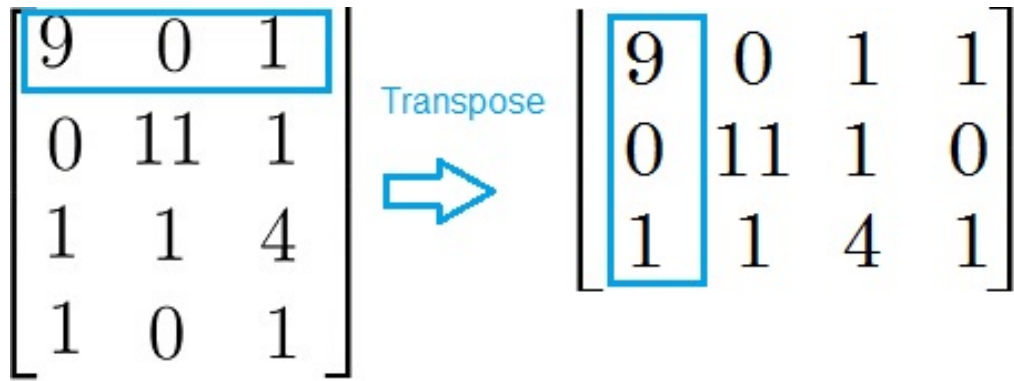

$$\begin{bmatrix} 9 & 0 & 1 \\ 0 & 11 & 1 \\ 1 & 1 & 4 \\ 1 & 0 & 1 \end{bmatrix} \xrightarrow{\text{Transpose}} \begin{bmatrix} 9 & 0 & 1 & 1 \\ 0 & 11 & 1 & 0 \\ 1 & 1 & 4 & 1 \end{bmatrix}$$

Figure 3: Operazione di trasposizione per chi non avesse ben capito. Altre informazioni e proprietà