Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea in

Statistica per le Tecnologie e le Scienze



La Rete della Fama: Network Analysis sui dati Pantheon

Relatore: prof.ssa Giovanna Menardi
Dipartimento di Scienze Statistiche

Laureando: Gabriele Paganelli
Matricola n. 2035052

Anno Accademico 2023/2024

Alla nonna Sandra,
che avrebbe meritato di essere
accanto a me per questo traguardo

Indice

Introduzione **1**

1 I dati **3**
1.1 I dati Pantheon . . . 3
1.1.1 Generalità . . . 3
1.1.2 La struttura di rete e le variabili a disposizione . . . 4
1.2 Caratteristiche delle figure storiche . . . 5
1.2.1 Le unità statistiche nello spazio e nel tempo . . . 5
1.2.2 Genere e impieghi delle unità statistiche . . . 6
1.2.3 Importanza delle unità statistiche . . . 7
1.3 Alcuni *pattern* multivariati . . . 8
1.3.1 Evoluzioni temporali . . . 8
1.3.2 Composizione dei domini . . . 10
1.3.3 Il numero di visite della pagina Wikipedia . . . 11
1.3.4 Comportamento di HPI . . . 13

2 Analisi esplorativa delle strutture di rete **15**
2.1 Introduzione alle reti . . . 15
2.1.1 Generalità . . . 15
2.1.2 Proprietà delle reti . . . 16
2.1.3 Quantità notevoli . . . 17
2.1.4 Le interazioni tra gruppi . . . 18
2.2 Gli indici di rilevanza . . . 19
2.2.1 Generalità . . . 19
2.2.2 Alcuni indici di rilevanza . . . 20
2.3 La rappresentazione grafica di una rete . . . 22
2.4 Analisi esplorativa della rete Pantheon . . . 24
2.4.1 Descrizione della rete globale . . . 24
2.4.2 Analisi per sottoreti . . . 24
2.4.3 I nodi più rilevanti . . . 31
2.5 Visualizzazione della rete Pantheon . . . 40
2.5.1 La rete intera . . . 40
2.5.2 Le sottoreti . . . 43

3 Analisi inferenziali delle strutture di rete **47**
3.1 Alcuni modelli di rete . . . 47
3.1.1 I modelli di Erdős-Rényi, *p1* e *p2* . . . 47

3.1.2 Modelli a blocchi e a variabili latenti 48
3.2 Gli ERGM 49
3.2.1 Generalità 49
3.2.2 Stima e bontà di adattamento 51
3.3 ERGM su alcune sottoreti di Pantheon 53
3.3.1 Premesse 53
3.3.2 Un modello di indipendenza diadica 54
3.3.3 La rete dei personaggi irlandesi 55
3.3.4 Una rete selezionata pseudo-casualmente 57

Conclusioni **63**

A Grafici aggiuntivi **65**

Ringraziamenti **77**

Introduzione

Quali sono le persone più influenti nella storia? Cosa determina le interazioni tra persone attraverso il tempo e lo spazio? A partire dai dati Pantheon (Yu *et al.*, 2016; Beytía e Schobin, 2018), si può provare a rispondere a queste e una varietà di altre domande. In questa relazione, ci si servirà di metodi, descrittivi e inferenziali, propri dell'analisi dei dati di rete (*network analysis*) su questo *dataset*; la *network analysis* rappresenta infatti uno strumento potente per decifrare sistemi complessi di relazioni.

Pantheon contiene informazioni biografiche su più di 11 000 figure celebri, dal passato fino ai giorni nostri, rappresentando un'ampia varietà di provenienze geografiche e di ambiti culturali; Beytía e Schobin hanno ampliato ulteriormente il *dataset* includendo le connessioni tra gli individui e permettendo quindi l'uso dei dati sotto forma di rete. La fonte principale è Wikipedia (wikipedia.org), la nota enciclopedia online ampiamente utilizzata in molteplici studi accademici e non. La ricchezza dei dati permette dunque di dare uno sguardo alla storia dell'umanità, e di individuare caratteristiche ricorrenti sia nei suoi protagonisti, che nelle interazioni tra di essi.

L'analisi dei dati Pantheon sotto forma di rete è guidata dalla pervasività di strutture complesse e reticolari nelle esperienze umane. Il mondo in cui viviamo è notoriamente interconnesso, che lo si voglia intendere in termini economici, sociali o tecnologici. Questo aspetto si riflette nel crescente interesse che la comunità scientifica, e in particolare quella statistica, rivolge alla *network analysis*. Se in termini generali una rete viene descritta come un insieme di persone od oggetti interconnessi, in statistica una rete è caratterizzata da un insieme di connessioni, denominate archi, tra delle unità, dette nodi (o attori), di varia natura. Vi è quindi un'importante distinzione rispetto alla statistica "classica": l'informazione riguardo le unità statistiche non è data (solo) dalle caratteristiche misurate su queste, ma dalle interazioni tra nodi e gruppi di nodi. Risulta quindi evidente la necessità di metodi e modelli alternativi, dal momento che cambiano tanto le domande d'interesse quanto l'impostazione formale dell'analisi.

L'esempio di rete più ovvio che si può proporre è dato dalle interazioni o dalle relazioni tra persone in un dato contesto; difatti, a partire da Moreno e Jennings (1938), la *network analysis* è stata strettamente legata alle scienze sociali, quali sociologia, psicologia sociale e antropologia sociale. L'analisi delle reti sociali non ha fatto che aumentare con l'avvento del Web 2.0 e dei cosiddetti, appunto, *social network*. È però importante non tralasciare la moltitudine di campi in cui si sfruttano le reti: in ambito biologico, ad esempio con lo studio delle interazioni tra varie aree del cervello, come in ambito tecnologico, con l'analisi dei collegamenti tra gli aeroporti di una data regione.

La flessibilità degli approcci basati sulle reti e la natura reticolare di una varietà

di fenomeni hanno quindi determinato la diffusione della *network analysis*, che a sua volta ha dato impulso alla ricerca. Nonostante i notevoli progressi degli ultimi 30 anni, vi sono però diversi ostacoli, sia interpretativi che algoritmici, che tengono a freno l'applicazione di molti interessanti risultati teorici su larga scala e su grandi moli di dati. Un problema con cui ben presto ci si scontra, ad esempio, è l'elevatissima complessità computazionale di alcuni algoritmi, sia per fini analitici che inferenziali. Difatti, anche per un numero di nodi limitato, ci si trova a gestire una quantità di possibili reti sostanzialmente infinita: ciò esclude un approccio enumerativo e richiede tecniche alternative anche per l'ottimizzazione. Anche questi problemi sono presentati nella presente relazione.

La trattazione e l'analisi della rete Pantheon si sviluppano come segue.

Nel Capitolo 1 si presentano le caratteristiche dei nodi della rete Pantheon, con l'analisi esplorativa delle principali variabili a disposizione. Verranno quindi descritti dei (potenziali) pattern, ed eventuali *bias* da tenere in considerazione.

Nel Capitolo 2 si affrontano le reti, i relativi indici di centralità e le difficoltà presentate dalla rappresentazione grafica dei dati di rete. Si esplorano possibili tendenze nelle distribuzioni degli archi e i nodi più centrali. Si evidenzia poi la concordanza tra le analisi svolte e alcune rappresentazioni grafiche della rete.

Nel Capitolo 3 viene affrontato il problema dell'inferenza, con la presentazione di alcuni modelli per dati di rete. In particolare, si approfondiscono le caratteristiche dei modelli della famiglia esponenziale per grafi casuali e i relativi vantaggi e svantaggi. Questi sono poi indagati tramite la stima di alcuni modelli, di cui si presentano anche le difficoltà computazionali e la bontà d'adattamento.

Nelle Conclusioni, si riassume il lavoro svolto, delineando i principali risultati ottenuti. Si evidenziano gli aspetti più significativi e le eventuali criticità emerse; sono poi suggeriti possibili strade alternative da esplorare per ampliare ulteriormente la comprensione della rete Pantheon.

Capitolo 1

I dati

1.1 I dati Pantheon

1.1.1 Generalità

Pantheon (versione 1.0, Yu *et al.* (2016)) è un insieme di dati che racchiude informazioni biografiche su 11 341 persone di rilevanza storica; la discriminante per l'inclusione di un soggetto nel *dataset* è che la sua biografia su Wikipedia compaia in più di 25 lingue diverse. La raccolta dati è stata svolta tra il 2012 e il 2013: mentre le informazioni biografiche, come il luogo o l'anno di nascita di un individuo, non sono influenzate da questo dettaglio, è ovvio come il numero di lingue in cui è scritta una data biografia o le visualizzazioni che essa riceve siano invece strettamente legati alla data di rilevazione. La soglia minima posta sul numero di lingue serve a trovare un compromesso per garantire che la popolarità dei personaggi selezionati non sia solo locale, ma abbia una copertura sufficientemente globale. Gli autori sostengono infatti che in questo modo la tendenza a includere personalità della cultura occidentale sia effettivamente ridotta.

Nel *dataset* è poi introdotta una tassonomia per classificare le unità statistiche in base al lavoro che hanno svolto: dal generale al particolare, si passa dal dominio (*domain*), all'industria (*industry*), e infine all'occupazione (*occupation*). Per la tassonomia completa e le scelte che hanno portato alla sua formazione, si rimanda all'articolo originale (Yu *et al.*, 2016). Un altro dato da tener presente riguarda il Paese di nascita degli individui: è stato scelto, per ragioni prevalentemente pratiche, di utilizzare i confini geografici moderni. Questa scelta rientra, assieme all'uso di Wikipedia come fonte, tra le principali fonti di *bias* nei dati, come evidenziato dagli autori stessi. Difatti, sono state osservate diverse tendenze dei redattori di articoli dell'enciclopedia *online*: mediamente occidentali e colti, gli *editors* tendono a dare più copertura a personaggi maschili, occidentali, anglofoni e nati di recente. Inoltre, Wikipedia presenta dei *bias* sistematici nella copertura di diversi domini d'interesse. Un approfondimento su quanto appena menzionato si può trovare Brown (2011).

Infine, un ultimo importante apporto di Yu *et al.* è l'introduzione dell'indice di popolarità storica, o HPI (*Historical Popularity Index*). Questo indice considera il numero di lingue di una biografia su Wikipedia, il numero di visite che riceve, la loro distribuzione nello spazio e nel tempo e il tempo passato dalla nascita dell'individuo in questione per provare a quantificare la popolarità delle unità statistiche.

1.1.2 La struttura di rete e le variabili a disposizione

Successivamente, Beytía e Schobin (2018) hanno arricchito il *dataset*, con l'aggiunta di connessioni tra le personalità storiche, ma anche con la rilevazione e l'aggiunta di dati biografici e metriche e statistiche di rete per ogni unità statistica. Per la rilevazione delle interazioni, si sono considerati gli *hyperlink* su Wikipedia tra una biografia in lingua inglese e l'altra (in data 16/04/2018). Ciò implica che le connessioni non siano necessariamente simmetriche all'interno del dataset. La scelta della biografia in lingua inglese per la raccolta degli archi, oltre a essere un'altra fonte di *bias*, ha causato l'esclusione di un'unità statistica: si tratta del fotografo italiano Augusto de Luca. Ciò porta il numero di nodi a 11 340. Beytía e Schobin introducono anche l'indice di centralità biografica, BCI (*Biographical Centrality Index*). Si tratta di un indice normalizzato, che tramite l'algoritmo *Pagerank* (si veda il Capitolo 2) e il numero di lingue di una biografia cerca di quantificare il grado di connettività e interculturalità della persona a cui si riferisce. Gli autori evidenziano come questo indice sia adatto più ad analizzare l'organizzazione delle informazioni in Wikipedia, che come *proxy* per la rilevanza storica degli individui; difatti, l'importanza dei personaggi del XX secolo risulta significativamente gonfiata, risultando nel cosiddetto *recency bias*. Gli autori concludono il *paper* con delle brevissime analisi e degli esempi di possibili domande a cui si può rispondere con i dati a disposizione.

Delle 42 colonne, e quindi variabili, presenti nei dati della rete Pantheon, alcune verranno ignorate nelle analisi di questa relazione. Certe perché riportano informazioni non d'interesse sulla pagina Wikipedia delle unità statistiche, altre perché riportano il valore di indici di centralità che saranno calcolati all'occorrenza nel corso delle analisi. Viene inoltre aggiunta la variabile *century*, calcolata come secolo di nascita delle singole unità. I dati utilizzati si compongono quindi di una matrice di dimensioni 126155×2, che descrive la lista delle connessioni (*edge list*), e di una matrice dei dati di dimensioni 11340×16, contenente per ogni personaggio storico le seguenti variabili:

- **Id**: codice identificativo unico per ogni unità statistica (numero naturale)
- **Name**: nome del personaggio storico (in inglese)
- **continentName**: continente natale dell'individuo (modalità: i 6 continenti e "Unknown", se sconosciuto)
- **countryName**: Paese natale dell'individuo (194 modalità, compreso "Unknown", se sconosciuto)
- **birthcity**: città natale dell'individuo (5093 modalità)
- **birthyear**: anno di nascita dell'individuo (compreso fra 3500 a.C e 2018 d.C)
- **deathyear**: anno di morte dell'individuo (2018 se l'individuo non è morto)
- **century**: secolo di nascita dell'individuo
- **agespan**: età dell'individuo alla morte (età corrente se è ancora in vita)
- **occupation**: occupazione dell'individuo (88 modalità)

- **industry**: categoria basata su un'aggregazione di occupazioni correlate (27 modalità)
- **domain**: categoria basata su un'aggregazione di industrie correlate (8 modalità: *Institutions*, *Public Figure*, *Arts*, *Sports*, *Science*, *Humanities*, *Business*, *Exploration*)
- **domain2**: come **domain**, ma la modalità *Institutions* si divide in *Army*, *Government* e *Religion*
- **gender**: il genere dell'individuo, maschio o femmina
- **AverageViews**: visualizzazioni medie della pagina Wikipedia per lingua (da gennaio 2008 a dicembre 2013)
- **HPI**: Indice di Popolarità Storica (equazione 4 in Yu *et al.*, 2016), varia da 9.88 a 31.99
- **BCI**: Indice di Centralità Biografica (Beytía e Schobin, 2018), normalizzato, quindi compreso tra 0 e 1

Per le analisi si è utilizzato il software `R` (R Core Team, 2023); i grafici sono stati realizzati con la libreria `ggplot2` (Wickham, 2016).

1.2 Caratteristiche delle figure storiche

1.2.1 Le unità statistiche nello spazio e nel tempo

Le unità statistiche del *dataset* presentano una certa eterogeneità sia nella provenienza che nel periodo storico in cui sono nate. Come è ragionevole aspettarsi, alcuni Paesi e alcuni secoli hanno una presenza molto più forte di altri, per ragioni storiche, sociali, demografiche o per i *bias* già discussi presenti nei dati.

Da un punto di vista geografico, analizzando le popolosità dei continenti si nota una grande presenza di personalità europee, che costituiscono più di metà del *dataset*, mentre oltre un quinto dello stesso è dato da individui nordamericani (Figura 1.1). Lo squilibrio nella rappresentazione dei continenti si riflette chiaramente anche nella distribuzione dei Paesi più rappresentati nei dati. Questa distribuzione segue un andamento esponenziale inverso, che si potrebbe anche interpretare come una legge di potenza. Gli Stati Uniti da soli costituiscono quasi un quinto del *dataset*, mentre Regno Unito, Francia, Italia e Germania insieme arrivano a quasi un terzo. Queste 5 nazioni sono le uniche ad avere una frequenza assoluta superiore a 500. Nei 15 Paesi più presenti non vi sono inoltre nazionalità sudamericane, africane o oceaniane, che sono i 3 continenti dalla presenza meno numerosa nei dati.

Un'analisi delle città di nascita più presenti è complicata dall'eterogeneità di questa variabile: vi sono ben 5093 diverse città; la modalità più presente è comunque "Other", che rappresenta tutte le città non menzionate esplicitamente, e ha una frequenza percentuale dell'11%. Tre quarti delle città compaiono una volta sola; anche qui risultano particolarmente rilevanti città situate in Europa e Nord America, in linea con i risultati ottenuti in precedenza.

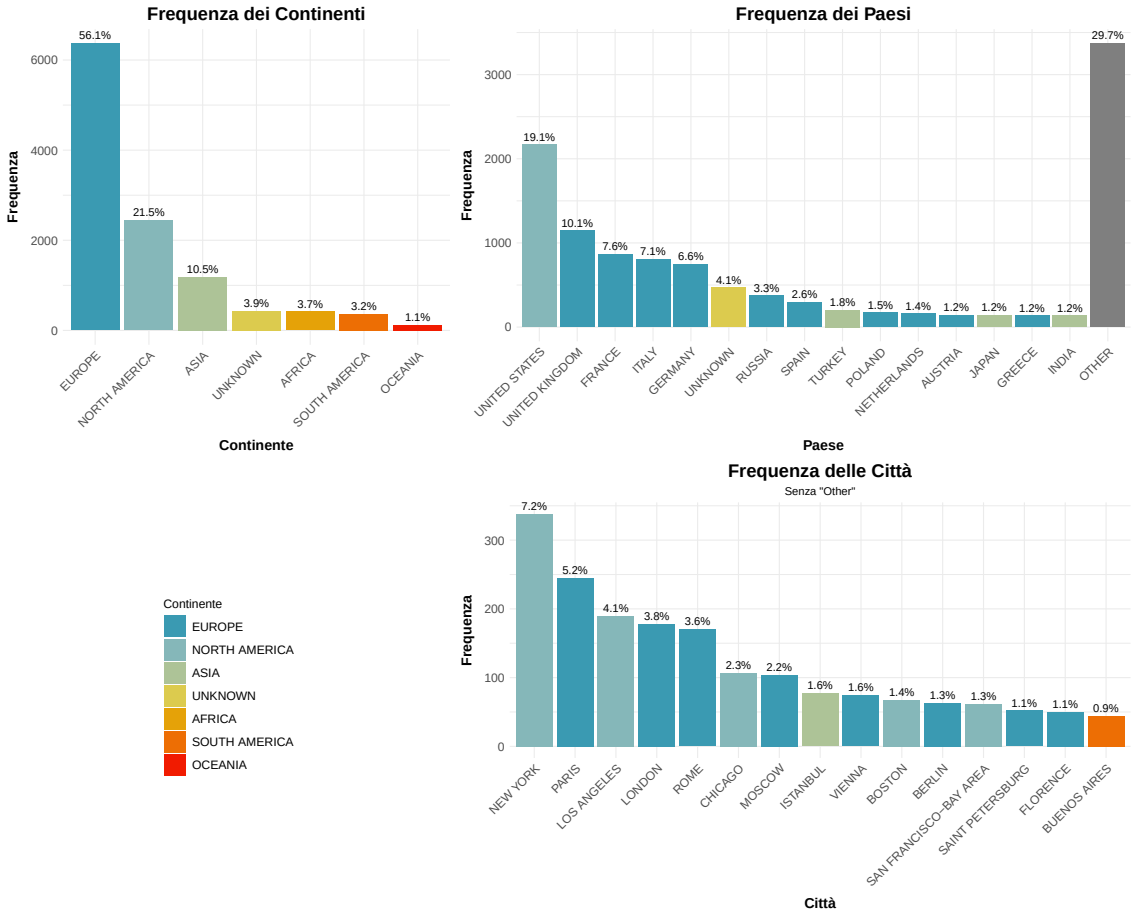


Figura 1.1: Frequenze relative e assolute di continenti, Paesi e città

In modo analogo alla distribuzione spaziale appena vista, la distribuzione delle unità statistiche nel tempo è piuttosto disomogenea (Figura 1.2). Ciò risulta evidente dalla crescita delle frequenze che si osserva nei secoli a noi più prossimi: l'andamento è approssimativamente esponenziale. Inoltre, i secoli antecedenti l'anno 0 hanno numerosità trascurabili se presi singolarmente, mentre solo i secoli XVIII, XIX e XX superano il mezzo migliaio di individui. Il XX secolo, da solo, costituisce più di metà del *dataset*, a conferma dei *bias* menzionati in precedenza.

Possiamo dunque affermare che la presenza di un individuo su Wikipedia di lingue diverse sembra favorita dall'appartenenza alla cultura occidentale e dalla vicinanza temporale al presente. Ciò non implica però che i personaggi più importanti, né nel *dataset* né nella storia, rispecchino queste caratteristiche demografiche.

1.2.2 Genere e impieghi delle unità statistiche

Gli stessi squilibri sottolineati per i luoghi e gli anni di nascita si presentano anche nel genere e negli impieghi associati alle varie personalità presenti nei dati.

La presenza di uomini è infatti abbondantemente maggiore rispetto a quella delle donne (Tabella 1.1), probabilmente a causa della disparità di genere che caratterizza la storia umana. Questo divario emerge dal fatto che per ogni donna siano presenti oltre 6 uomini.

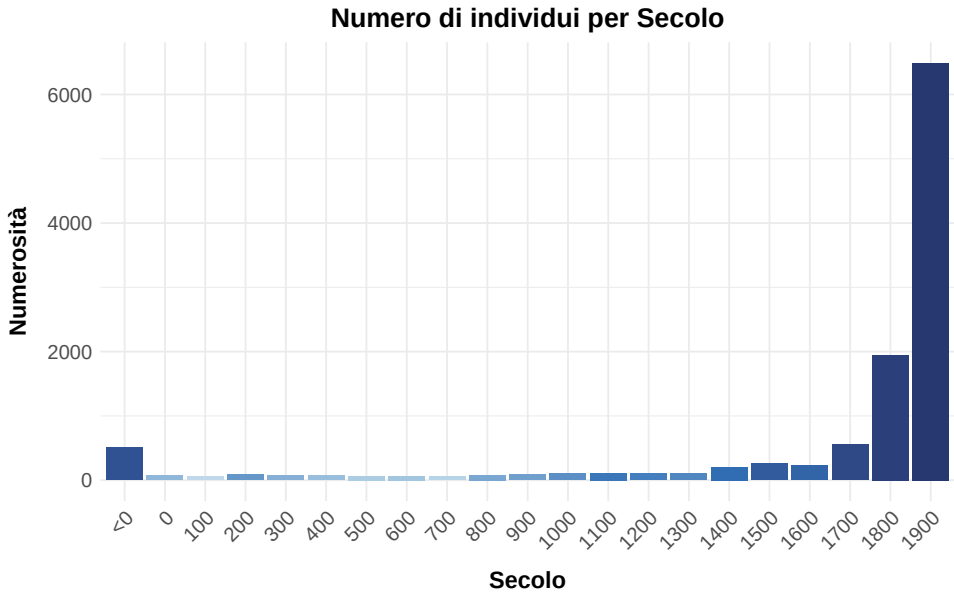


Figura 1.2: Numerosità di soggetti per ogni secolo

	Freq. Ass.	**Freq. %**
Donne	1495	13.2%
Uomini	9845	86.8%

Tabella 1.1: Frequenze assolute e percentuali dei generi

È d'interesse anche osservare a quali "domini", intesi come ambiti lavorativi, appartengono le occupazioni associate agli individui. La maggior parte del *dataset* è costituito da figure istituzionali o artistiche, ma non sono trascurabili sportivi, scienziati e umanisti (Figura 1.3). Un'analisi più approfondita, che prende in considerazione le occupazioni stesse, evidenzia come oltre metà dei dati riguardino politici, artisti, calciatori o scrittori; la distribuzione delle frequenze dei vari mestieri rimanda ancora a un andamento esponenziale negativo, dove poche modalità hanno frequenze alte e molte modalità hanno frequenze basse. Inoltre, le occupazioni principali appartengono ai domini più presenti tra le unità statistiche, come prevedibile. Sembra quindi che impieghi in alcuni ambiti favoriscano la popolarità dei personaggi.

1.2.3 Importanza delle unità statistiche

Il numero di visualizzazioni medie della pagina di un personaggio non sembra un buon indicatore della sua importanza storica, data l'elevata volatilità di questo valore e l'influenza che eventi temporanei possono avere sul dato. Difatti, tra le persone le cui pagine Wikipedia ricevono più visite troviamo figure popolari principalmente nella cultura nordamericana (Tabella 1.2): ad esempio, le pagine più visitate sono quelle di Kim Kardashian, Lil Wayne ed Eminem. Nonostante la loro innegabile influenza culturale odierna, è plausibile ritenere che nei 5500 anni coperti dai dati Pantheon vi siano personaggi che hanno contribuito di più alla storia dell'umanità.

È da questo presupposto che nasce l'indice HPI (si veda la Sezione 1.1.1). Seppur imperfetto, come tutti gli indici, sembra riassumere in modo più soddisfacente

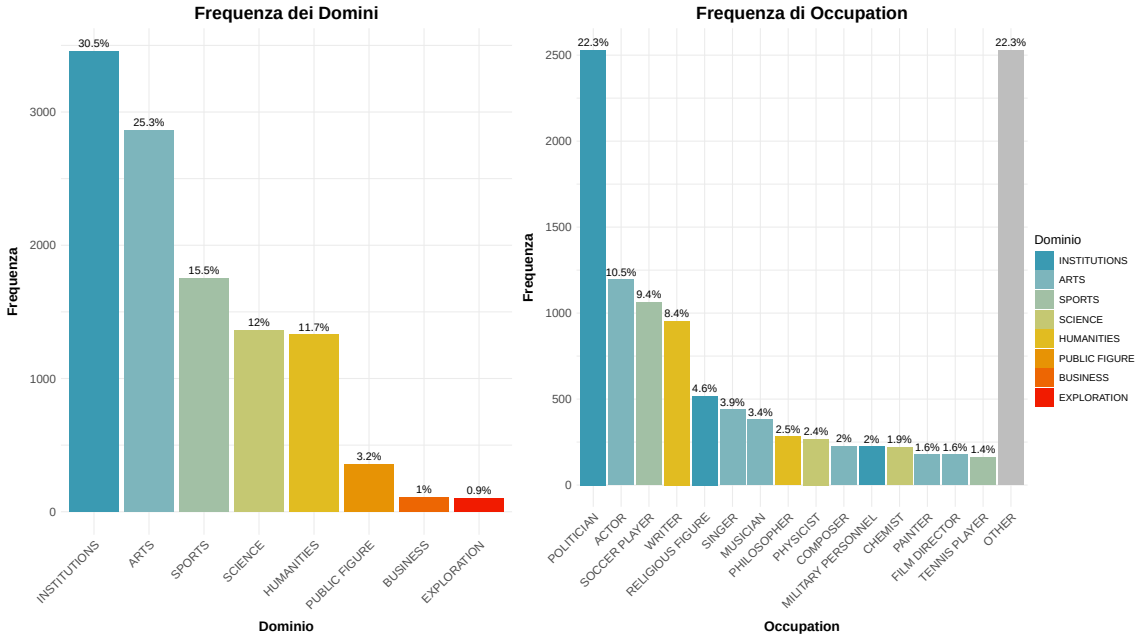


Figura 1.3: Frequenza assoluta e percentuale dei domini e nelle occupazioni

Individuo	**Views medie**	**Individuo**	**HPI**
Kim Kardashian	1 515 232	Aristotle	31.99
Lil Wayne	1 401 837	Plato	31.99
Eminem	1 312 695	Jesus Christ	31.90
Miley Cyrus	1 277 052	Socrates	31.65
Justin Bieber	1 208 065	Alexander the Great	31.58
Nicki Minaj	1 115 701	Leonardo da Vinci	31.46
The Rock	1 101 973	Confucius	31.37
Sasha Grey	1 061 930	Julius Caesar	31.12
Rihanna	1 058 112	Homer	31.11
Cristiano Ronaldo	1 053 770	Pythagoras	31.07

Tabella 1.2: Individui più rilevanti per views e HPI

l'importanza dei personaggi storici, tenendo conto di quanto sia diffusa la popolarità di un individuo nello spazio e nel tempo e della distanza temporale dai giorni nostri. In questo caso, i personaggi principali risultano essere Aristotele, Platone e Gesù Cristo, e in generale gli individui con un indice più elevato sono figure storiche di rilievo e dalla popolarità interculturale. Sarà interessante confrontare questi risultati con quelli ottenuti dai vari indici che la *network analysis* ci mette a disposizione.

1.3 Alcuni *pattern* multivariati

1.3.1 Evoluzioni temporali

Come si è visto, vi è una crescita esponenziale nel tempo delle unità statistiche incluse nei dati; la crescita non è però uniforme tra i continenti, e anzi alcuni presentano una decrescita, tanto in valore assoluto quanto percentuale (Figura 1.4). Le personalità

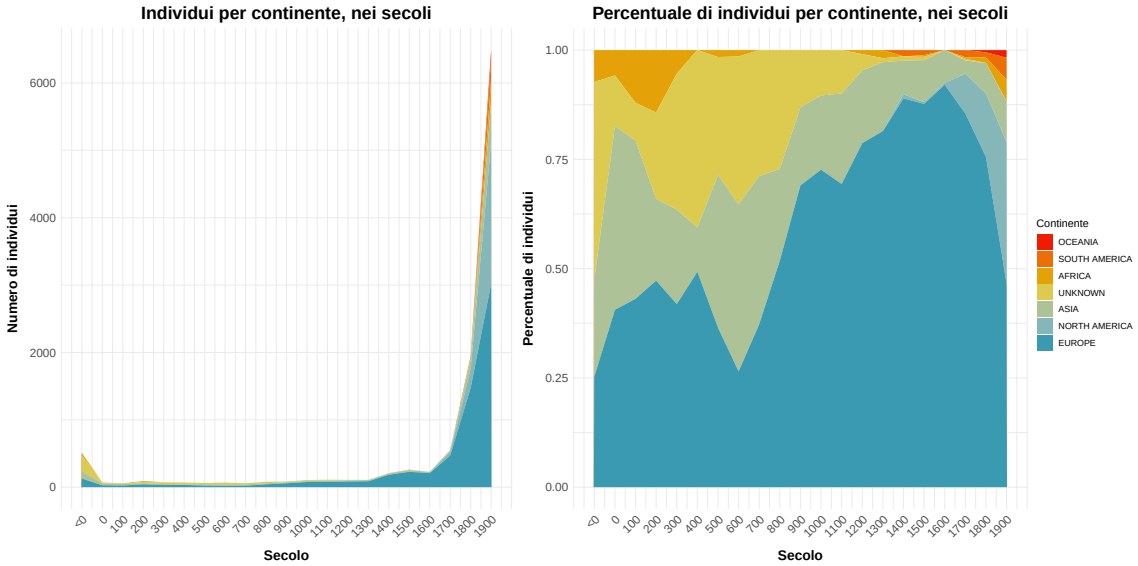


Figura 1.4: Frequenza assoluta e percentuale dei continenti nei secoli

europee hanno una certa rilevanza in ogni periodo storico, e arrivano a rappresentare la stragrande maggioranza dei dati intorno al 1600, con il Rinascimento; nei secoli successivi, gli altri continenti tornano a essere più rappresentati. Inizialmente non trascurabile, con il tempo diminuisce la rilevanza di personaggi africani, asiatici, o il cui continente di provenienza è sconosciuto. In quest'ultimo caso si può attribuire il calo alla migliore informazione a disposizione, e quindi alla diminuzione di individui con provenienza sconosciuta; è invece noto come Africa e Asia, nel tempo, abbiano presentato meno figure popolari, almeno nella cultura occidentale. Infine, nonostante la completa assenza fino al XVI secolo, le figure nordamericane coprono una porzione importante dei dati nei secoli più recenti, coerentemente con la rilevanza assunta dal continente dopo la sua "scoperta" avvenuta nel 1492.

Similmente, il disequilibrio tra uomini e donne è variabile nei secoli (Figura 1.5). In particolare, c'è un leggero ma progressivo aumento della percentuale di donne fino al 1800, momento in cui questa crescita sembra accelerare. Inoltre, lo stesso fenomeno di lenta ma costante crescita sembra verificarsi nel XX secolo, dove gli anni più recenti tendono a presentare una proporzione di uomini e donne più bilanciata, coerentemente con la considerazione, crescente ma tuttora deficitaria, della società verso le donne negli anni. Nonostante non si sia ancora arrivati a una vera e propria parità, sarebbe interessante svolgere queste analisi su dati più aggiornati, sia per verificare se la maggior attenzione alle pari opportunità degli ultimi anni abbia avuto influenza sull'organizzazione di Wikipedia, sia per vedere se le figure di rilievo più recenti presentino un effettivo equilibrio di genere.

A variare nel tempo, è abbastanza prevedibile, sono anche i domini d'impiego delle unità statistiche (Figura 1.6). I secoli antecedenti l'anno 1000 vedono la presenza quasi unicamente di figure istituzionali e umanistiche; subito dopo, aumenta la presenza di personalità legate all'arte. Il settore dell'esplorazione è rappresentato in maniera significativa solo negli anni intorno al 1500, con la già menzionata scoperta dell'America; negli stessi anni, le figure scientifiche vedono crescere la propria numerosità, soprattutto nel periodo dell'Illuminismo. Infine, il mondo dello

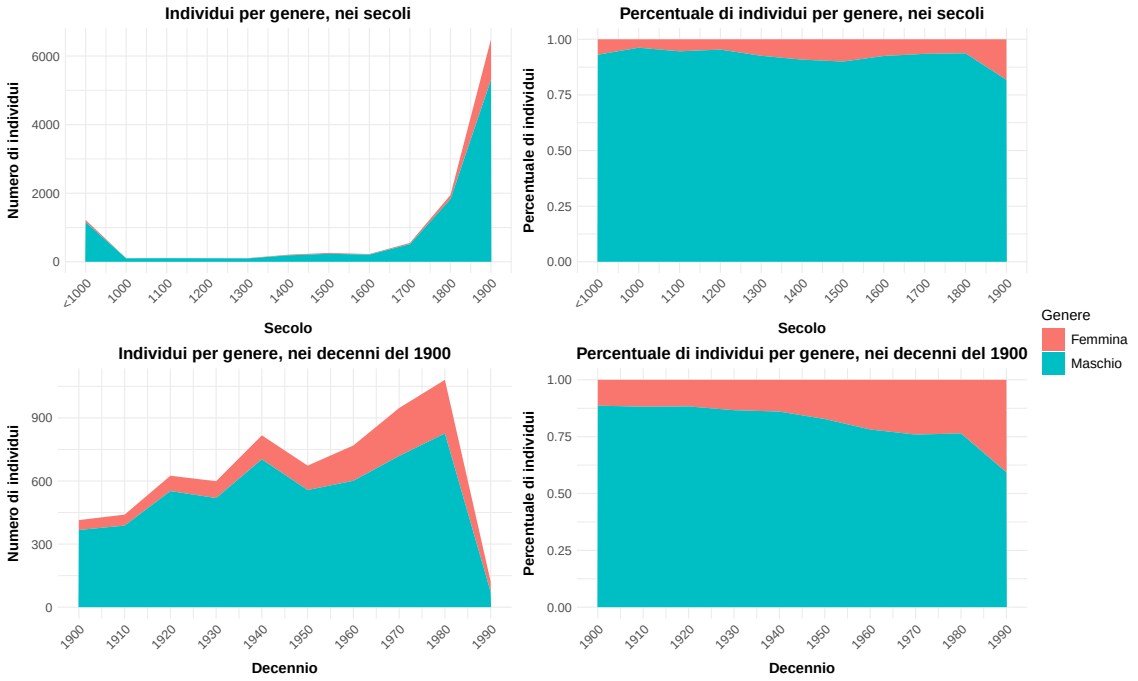


Figura 1.5: Frequenza assoluta e percentuale dei generi nei secoli e nel Novecento

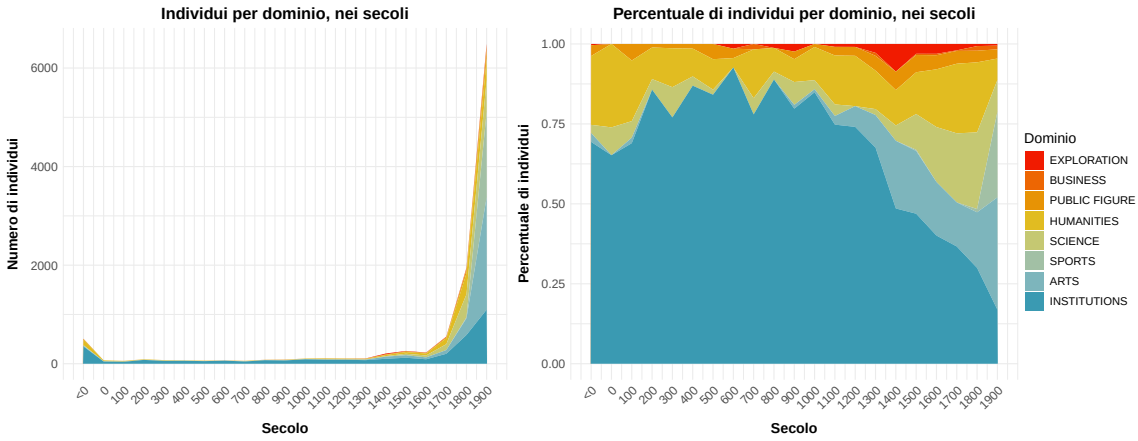


Figura 1.6: Frequenza assoluta e percentuale dei domini nei secoli

sport è completamente assente prima del 1900, ma in questo secolo la sua presenza è notevole, seconda solo alle arti e addirittura superiore alle istituzioni. In generale, si può dire che i personaggi che hanno ricoperto ruoli istituzionali sono quelli con la presenza maggiore nel corso del tempo, ma negli anni più recenti c'è un maggior equilibrio nella diffusione dei domini nel *dataset*.

1.3.2 Composizione dei domini

Ad avere un'influenza sull'ambito d'interesse di un dato individuo non è solo il periodo storico in cui vive, ma anche il luogo di nascita. Difatti, i diversi domini presentano grandi differenze nei Paesi più frequenti al proprio interno; ci sono alcuni domini in cui la maggior parte della popolazione proviene da pochi Paesi, e altri che invece presentano delle distribuzioni più variegate (Tabella 1.3). Ad esempio, il campo

INSTITUTIONS	Perc.	**ARTS**	Perc.	**SPORTS**	Perc.
Unknown	12%	United States	41%	United Kingdom	9%
Italy	10%	United Kingdom	14%	United States	9%
France	6%	France	7%	Germany	7%
Germany	6%	Italy	6%	Italy	7%
United Kingdom	6%	Germany	4%	Spain	6%
SCIENCE	Perc.	**HUMANITIES**	Perc.	**OTHER**	Perc.
United States	25%	France	13%	United States	27%
United Kingdom	14%	United States	12%	United Kingdom	11%
Germany	12%	United Kingdom	10%	France	7%
France	10%	Germany	8%	Germany	6%
Russia	4%	Italy	7%	Russia	4%

Tabella 1.3: Paesi più rappresentati in ogni dominio e relativa percentuale

dell'arte è popolato in gran parte da statunitensi e britannici, probabilmente anche grazie alla popolarità del cinema e di Hollywood; al contrario, gli sport sembrano presentare una distribuzione di provenienza molto più equa. Nelle istituzioni, la provenienza sconosciuta è la più frequente: è plausibile che sia dovuto alle poche informazioni a disposizione su personaggi molto antichi, che sono tendenzialmente figure politiche. L'Italia è la seconda nazionalità più presente in campo istituzionale; ciò è probabilmente spiegabile grazie alla tradizione cattolica e a tutte le figure a essa legate, quali papi e vescovi.

Nei vari domini la presenza di donne non risulta costante (Tabella 1.4): campi considerevolmente squilibrati a sfavore del genere femminile sono le scienze, gli sport e l'ambito istituzionale. L'arte presenta più di metà delle donne di tutto il *dataset*; in generale, solo in ambito artistico e di personalità pubbliche le donne sono mediamente più presenti degli uomini. Anche in questo caso, quindi, i dati Pantheon sembrano riflettere la storia umana e la società odierna, con le sue disuguaglianze e i suoi *bias*.

L'analisi della distribuzione dei generi nei domini, condizionata all'anno di nascita, non rivela molto altro (Figura 1.7): la maggior parte delle (poche) donne presenti nel *dataset* e nate prima del 1900 sono figure pubbliche, istituzionali o umanistiche. Gli uomini sono invece principalmente attivi nel campo di istituzioni, lettere, scienze e arte. Nel XX secolo le donne diventano molto presenti in campo artistico, con una numerosità non trascurabile anche nello sport, nelle istituzioni e nelle figure pubbliche; i settori più popolati da personalità maschili diventano lo sport e l'arte, con una presenza sostanziale anche in istituzioni e scienze. In generale, si conclude che l'unico campo dove la presenza femminile risulta corposa è l'arte, e tale presenza si concentra nel XX secolo. Questa conclusione non è sorprendente: quasi metà delle donne nel *dataset* sono attrici o cantanti, impieghi che appartengono al mondo artistico.

1.3.3 Il numero di visite della pagina Wikipedia

Abbiamo già visto come le visite (*views*) alla pagina Wikipedia di un personaggio non sembrino un *proxy* affidabile della sua rilevanza storica, ma solo di quella attuale, e in particolare nella cultura nordamericana. Uno sguardo alle *views* medie e totali

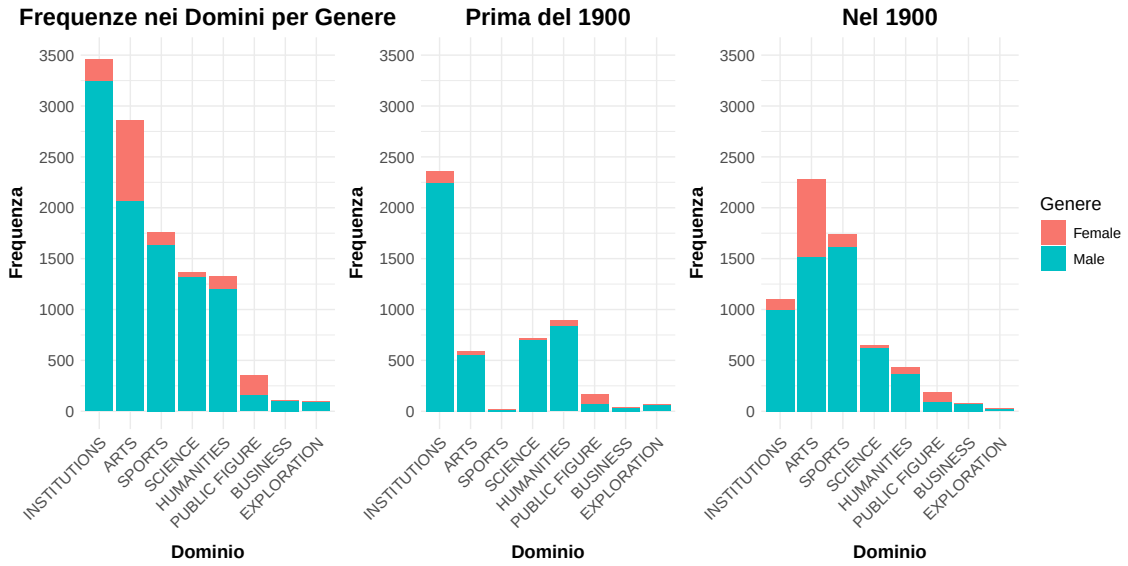


Figura 1.7: Frequenze assolute dei generi nei domini, anche condizionate al secolo

Totale	Science	Business	Institutions	Exploration
6.6	30.8	20.6	15.2	13.6
	Sports	Humanities	Arts	Public Figure
	13.5	10.1	2.6	0.9

Tabella 1.4: Rapporto uomini/donne totale e per dominio

(Figura 1.8) conferma che i personaggi nordamericani hanno il maggior numero di visualizzazioni, sia totali che medie. Gli europei sono secondi per visualizzazioni totali, ma le personalità provenienti da Sud America e Oceania hanno mediamente più *views*: ciò è probabilmente dovuto alla scarsità di sudamericani e oceaniani nel *dataset*, che gonfia i valori medi.

Statunitensi e canadesi hanno i valori medi di visite più alti; i primi hanno anche il totale di visualizzazioni di gran lunga maggiore, mentre i secondi, complice una numerosità ridotta, non spiccano in questa statistica. Mentre la distribuzione delle *views* totali sembra seguire approssimativamente quella delle numerosità nel *dataset*, sembra che il valore medio sia invece influenzato dalle caratteristiche demografiche odierne dei Paesi in questione: Giappone, India e Brasile sono infatti particolarmente popolosi e ciò si riflette nell'elevato numero di visite medie alle pagine Wikipedia di persone nate in tali nazioni.

L'andamento delle *views*, sia medie che totali, è poco sorprendente al variare dei secoli: il I secolo, e gli anni antecedenti, hanno un numero medio di visualizzazioni elevato, principalmente grazie alla presenza di figure legate alla cristianità; gli anni successivi non presentano valori notevolmente elevati. Del resto, avvicinandosi al presente, aumentano mediamente sia le visualizzazioni medie che quelle totali. La grande numerosità di persone nate nel XX secolo e la tendenza delle persone a cercare su Wikipedia i propri contemporanei, fanno sì che questo secolo abbia di gran lunga il maggior numero di *views* sia totale che medio.

Per quanto riguarda i domini, invece, le figure del settore artistico sono quelle che attirano indubbiamente più visualizzazioni. L'andamento delle *views* medie

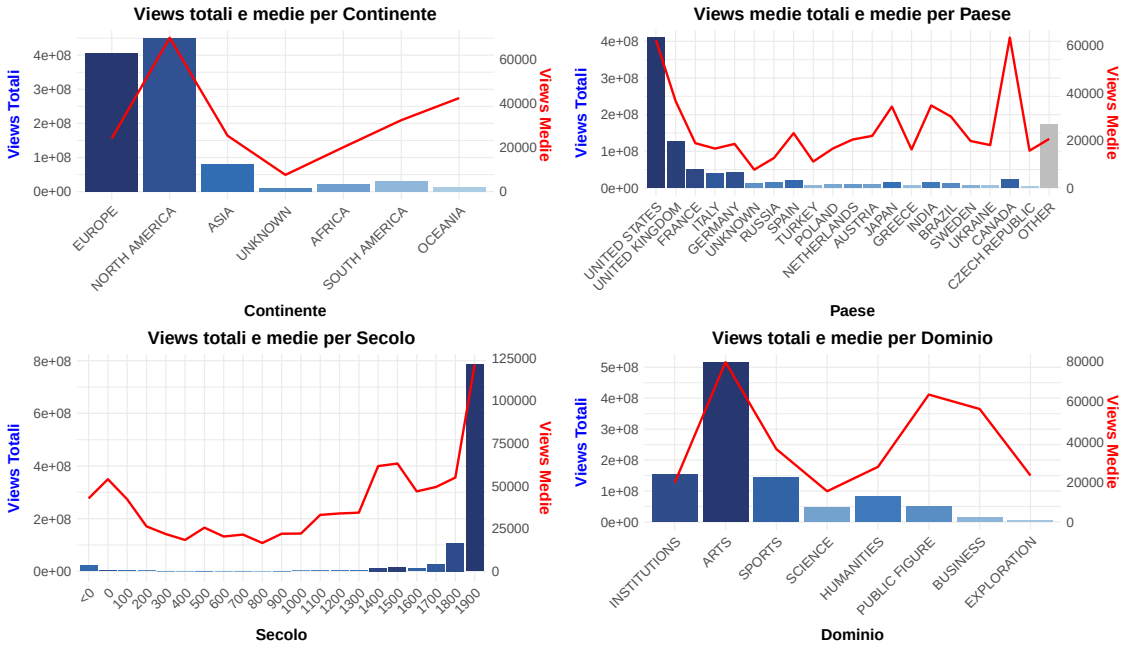


Figura 1.8: Views medie e totali per continente, Paese, secolo e dominio

all'interno dei domini è caratterizzato da una certa coerenza con quello delle *views* totali: gli ambiti con più visite medie hanno anche più visite cumulate, e questi dati sono spesso analoghi alle popolosità dei vari domini nel *dataset*. Rappresentano delle eccezioni le personalità pubbliche o del mondo degli affari; la loro modesta numerosità causa un basso numero di *views* totali, ma la loro popolarità fa sì che il valore medio sia molto elevato.

1.3.4 Comportamento di HPI

Come osservato in precedenza, l'indice HPI sembra individuare le personalità più importanti nel *dataset* in maniera, se non affidabile, quantomeno in linea con il senso comune. L'analisi dei continenti, Paesi, secoli o domini che hanno cumulativamente un indice HPI più elevato non presenta sorprese (Figura 1.9): i gruppi di unità statistiche più popolosi hanno anche il valore totale di HPI più alto. Ad esempio, Stati Uniti e 1900 sono rispettivamente il Paese e il secolo più presenti nel *dataset*, e presentano anche il totale di popolarità storica maggiore.

Molto interessante è uno sguardo al valore medio dell'HPI: questo indice ha infatti un comportamento estremamente stabile. Sia al variare del continente, che al variare del Paese, l'indice ha un valore medio tra il 20 e il 25, nonostante il *range* osservato nei dati vari in modo ampio, tra 9 e 32. Risultano leggermente più importanti Paesi più rilevanti nell'antichità, come Grecia o Turchia, ma non sembrano esserci classi di unità statistiche che l'HPI mette in risalto.

Più curioso è il comportamento medio dell'indice nel tempo: avvicinandosi ai giorni nostri, i secoli hanno un'importanza che pare lievemente decrescente, per poi calare considerevolmente nel 1900. Ciò si può attribuire presenza di pochi individui, ma importanti, nei secoli più antichi; inoltre, la formula dell'HPI cerca di valorizzare i personaggi nati in tempi più antichi, per bilanciare il *recency bias*, e pare ci riesca.

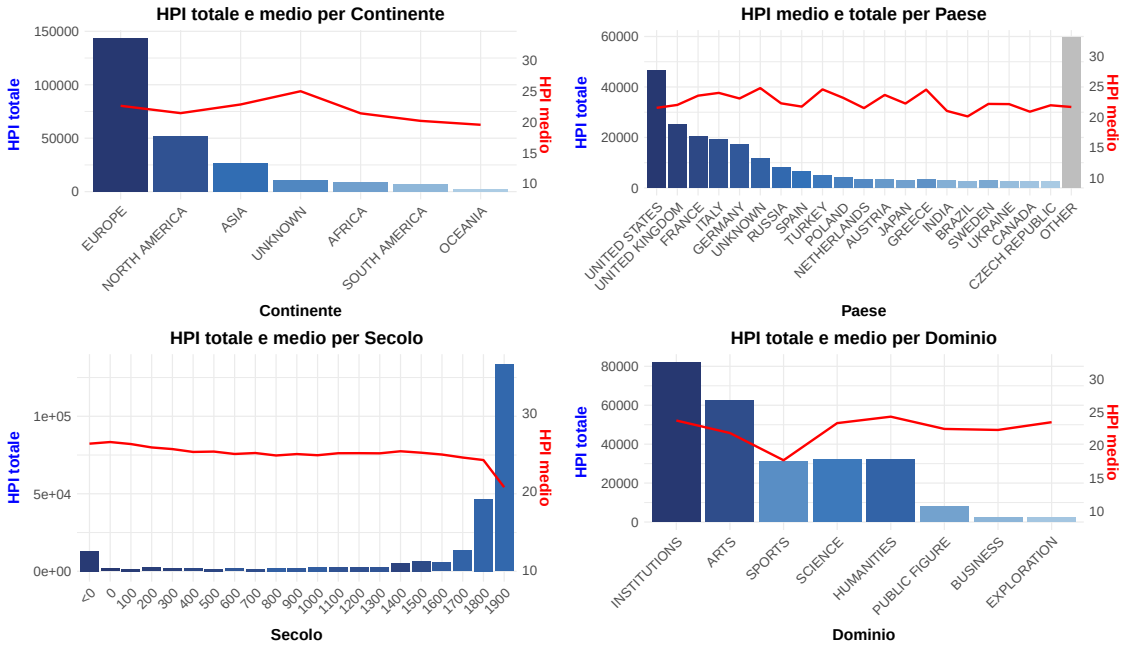


Figura 1.9: HPI medio e totale per continente, Paese, secolo e dominio

Anzi, si potrebbe argomentare che la "penalità" assegnata alle personalità più recenti rischi di essere fin troppo severa.

Altrettanto interessante è l'andamento dell'HPI medio nei domini: mantiene la sua caratteristica stabilità per tutti i maggiori ambiti, eccezion fatta per lo sport. Questo dominio ha una popolarità storica media particolarmente bassa. Si individuano due principali ragioni per questo: la prima è che tutte le unità statistiche a esso appartenenti sono nate nel XX secolo, e probabilmente soffrono del *recency bias* inverso che pare presentare l'indice; la seconda è che l'influenza della maggior parte degli sportivi risulta locale e quindi ridotta rispetto agli altri individui che fanno parte di Pantheon.

Capitolo 2

Analisi esplorativa delle strutture di rete

2.1 Introduzione alle reti

2.1.1 Generalità

Una rete è essenzialmente composta da nodi e archi. I nodi sono le unità statistiche su cui sono misurati due tipi di variabili: le variabili strutturali, ovvero gli archi, che definiscono le interazioni tra i nodi, e le variabili di composizione, o attributi, che sono invece rilevazioni di determinate variabili sui singoli nodi. Le reti sono generalmente definite servendosi del linguaggio della teoria dei grafi: un grafo $\mathcal{G}(\mathcal{N}, \mathcal{L})$ consiste in un insieme di N nodi $\mathcal{N} = \{n_1, ..., n_N\}$ e un insieme di L archi $\mathcal{L} = \{l_1, \ldots, l_L\}$. La cardinalità dal primo insieme determina l'ordine della rete, mentre quella del secondo insieme ne definisce la dimensione. Due nodi n_i e n_j collegati da un arco si dicono adiacenti.

Se gli N nodi della rete rappresentano attori dello stesso tipo si dice che la rete è monomodale (*one-mode*); questo è il caso di una semplice rete di amicizie, in cui i nodi sono tutti persone. Al contrario, una rete bipartita (*two-mode*) è costituita da nodi di due diverse entità, in cui le interazioni si verificano solo tra i due gruppi, ma non all'interno degli stessi (si veda Wasserman e Faust, 1994): un esempio può essere una rete che rappresenta le interazioni tra aeroporti e compagnie aeree, ove quindi non è definita un'interazione tra due aeroporti o tra due compagnie aeree.

Gli L archi possiedono due caratteristiche fondamentali. La prima è la direzionalità: essi possono infatti essere simmetrici (indiretti) o direzionati (diretti). Nel primo caso, il fatto che il nodo n_i sia collegato al nodo n_j implica anche l'opposto. Per il caso direzionato, è invece possibile che il nodo n_i sia collegato al nodo n_j, ma non sia vero il contrario. L'altra importante proprietà è il peso: gli archi possono essere binari o pesati. Nel primo caso si misura la sola presenza o assenza di un'interazione tra i nodi; gli archi pesati invece rappresentano l'intensità di una data interazione. Ovviamente il significato di direzionalità e peso degli archi varia da un contesto all'altro. Vi sono inoltre contesti in cui è sensato assumere che un nodo interagisca con se stesso, da cui la presenza di archi riflessivi nella rete.

Una possibile rappresentazione di una rete è definita dalla matrice di adiacenza $\mathbf{Y}$, di dimensioni $N \times N$, le cui entrate y_{ij} sono date dal valore dell'arco dal nodo n_i

al nodo n_j. Se l'arco non è presente, $y_{ij} = 0$; se la rete è binaria e l'arco è presente, $y_{ij} = 1$; nel caso di archi pesati, y_{ij} ha il valore del peso dell'arco in questione. Una rete con archi indiretti è rappresentata da una matrice simmetrica, poiché $y_{ij} = y_{ji}$; inoltre, in una rete senza nodi riflessivi $y_{ii} = 0$, ovvero $\mathbf{Y}$ ha solo valori nulli lungo la diagonale. Un'alternativa computazionalmente più gestibile per reti di grandi dimensioni è la creazione della lista degli archi (*edge list*): si tratta di una matrice E, in cui per l'arco l troviamo in e_{l1} il nodo sorgente e in e_{l2} il nodo bersaglio; eventualmente in e_{l3} è riportato il peso dell'arco.

Una rete può presentare nodi isolati, ovvero senza archi né entranti né uscenti, ed è bene tenerne conto poiché questi attori possono alterare il comportamento di alcuni algoritmi, in un modo analogo agli *outlier* in statistica classica. Similmente, sono di particolare interesse le cosiddette componenti connesse della rete: si tratta di sottoreti in cui ogni nodo è raggiungibile a partire da qualunque altro nodo della componente connessa in questione. Se la rete è diretta, la connessione può essere forte o debole. Nel primo caso vengono tenute in considerazione le direzioni degli archi quando li si percorre, mentre nel secondo caso ci si comporta come se la rete fosse indiretta. L'individuazione di nodi isolati e componenti connesse torna particolarmente utile, sia nel calcolo di alcuni indici che nella rappresentazione grafica delle reti.

2.1.2 Proprietà delle reti

Le reti presentano spesso alcune proprietà a livello macroscopico. Tra queste troviamo la presenza di comunità (*communities*), ovvero di gruppi con molte connessioni interne ma scarse verso altri gruppi. La formazione delle comunità è spesso, anche se non necessariamente, legata all'omofilia: si tratta della propensione di nodi simili a legarsi l'uno all'altro (*love the same;* Mcpherson *et al.*, 2001). È importante però tenere a mente che le comunità di una rete sono un concetto che si basa sugli archi e sulle adiacenze dei nodi, quindi su interazioni effettivamente osservate. L'omofilia si riferisce invece agli attributi dei nodi e a una tendenza che agisce in modo latente, più o meno spiccato, aumentando la probabilità che due nodi simili siano adiacenti. L'individuazione delle comunità, sostanzialmente un'applicazione del *clustering* alle reti, costituisce un campo di ricerca noto come *community detection*.

Nei contesti delle reti sociali l'omofilia è di particolare interesse, dato il suo stretto legame con il fenomeno della camera dell'eco (*echo chamber*). Con questo termine ci si riferisce ad ambienti in cui una persona si confronta solo con opinioni e informazioni che rafforzano le sue idee, in un classico esempio di *confirmation bias*. Gli effetti delle *echo chamber* sembrano particolarmente rilevanti nello scambio di opinioni politiche e di disinformazione. Alcuni studi suggeriscono che questo fenomeno favorisca la polarizzazione e l'estremismo, mentre altri sono più cauti e ridimensionano l'impatto reale delle *echo chamber*. Una copertura più completa dell'argomento è offerta da Kitchens *et al.* (2020) e riferimenti.

Spesso le reti presentano anche la proprietà "piccolo mondo" (*small world*), popolarmente conosciuta come "regola dei sei gradi di separazione": secondo questa teoria due persone qualunque sono separate da non più di 6 intermediari. Introdotte da Milgram (1967), e poi formalizzate quantitativamente da Watts e Strogatz (1998), nelle reti *small-world* è possibile raggiungere qualsiasi nodo da qualunque altro in pochi passi. Questa proprietà è stata osservata frequentemente e in una varietà di

ambiti, al punto da far avanzare l'ipotesi che sia una proprietà "quasi universale", e conseguentemente di dubbia utilità. In Lovekar *et al.* (2021) si trova un'analisi critica estesa e approfondita, affiancata da un approccio basato su test d'ipotesi.

Infine, l'invarianza di scala (*scale-free*) è la proprietà secondo cui un nodo nuovo nella rete tende a connettersi con i nodi più connessi (Barabási e Albert, 1999), evidenziando dunque una tendenza del "ricco" a diventare più ricco (*rich gets richer*). La distribuzione delle connessioni dei nodi tende quindi a seguire un andamento esponenziale negativo, e in particolare analogo alla legge di potenza; ciò risulta nella presenza di pochi nodi molto connessi (detti *hub*), che fungono spesso da ponte nel piccolo mondo della rete. Anche qui, vi è un dibattito sulla reale diffusione delle reti *scale-free*; si rimanda quindi a Holme (2019) e relativi riferimenti.

2.1.3 Quantità notevoli

Da un punto di vista quantitativo, una rete può essere descritta mediante il calcolo di alcuni indicatori.

In primo luogo, la densità (*density*) indica la tendenza generale degli attori di una rete a connettersi. È data dal rapporto tra il numero di archi osservati e quelli possibili; in una rete diretta, assumendo l'assenza di riflessività:

$$\mu = \frac{L}{N(N-1)} \tag{2.1}$$

La reciprocità (*reciprocity*) indica invece la tendenza di un nodo, in una rete diretta, a rendere reciproco un arco in esso entrante. Vi sono due modi per calcolare questo indice. Nel primo caso, si calcola semplicemente il rapporto tra gli archi ricambiati sul totale di archi nella rete. Questa misura si interpreta come un mero rapporto che viene osservato. Nel secondo caso, utilizzando la notazione definita dalla matrice di adiacenza,

$$\text{rec} = 1 - \frac{\sum_{i,j} |y_{ij} - y_{ji}|}{2 \sum_{i,j} y_{ij}}. \tag{2.2}$$

Questa quantità è interpretabile come una stima della probabilità che, dato un arco y_{ij}, sia presente anche la sua controparte y_{ji}; ciò la rende preferibile alla prima formula, dal momento che ha delle implicazioni non solo descrittive ma anche inferenziali.

Un altro concetto utile da esplorare in una rete è la transitività (*transitivity*). Parliamo di transitività quando, dati i nodi n_i, n_j e n_k, la connessione tra n_i e n_j e quella tra n_j e n_k implicano una probabilità più elevata di connessione tra n_i e n_k. Più formalmente, in una rete binaria:

$$\Pr\left(Y_{ik} \mid y_{ij} = y_{jk} = 1\right) \geq \Pr\left(Y_{ik} \mid y_{ij} = y_{jk} = 0\right)$$

Non c'è una definizione quantitativa univoca per la transitività di una rete, soprattutto nel caso diretto; vengono utilizzate diverse misure, che possono coinvolgere il conteggio dei "triangoli" osservati nella rete, spesso rapportato alle terne possibili o quelle connesse. Frequentemente viene usata una misura, anche chiamata coefficiente

di *clustering*, data dal numero di terne "chiuse" sul numero di terne totali; la sua definizione è immediata per le reti simmetriche, non altrettanto per quelle dirette.

Infine, è utile anche una quantità per misurare l'omofilia, relativamente a una variabile categoriale con k livelli. Ci serviamo dunque dell'assortatività nominale:

$$r = \frac{\sum_i^k e_{ii} - \sum_i^k a_i b_i}{1 - \sum_i^k a_i b_i} \tag{2.3}$$

dove e_{ii} è la frazione di collegamenti che connettono nodi dello stesso tipo i, a_i è la frazione di collegamenti che partono da nodi di tipo i e b_i è la frazione di collegamenti che arrivano a nodi di tipo i.

Si nota come questa misura quantifichi la tendenza dei nodi a connettersi con altri nodi che condividono la stessa categoria. Il coefficiente è normalizzato tra -1 e 1: un valore positivo indica presenza di omofilia, mentre un valore negativo rileva il fenomeno opposto (disassortatività); un coefficiente nullo segnala connessioni casuali rispetto alle categorie.

2.1.4 Le interazioni tra gruppi

Dopo aver analizzato alcune caratteristiche globali della rete, è d'interesse comprendere anche come interagiscono diversi gruppi di attori, o sottoreti, al suo interno, definiti dai nodi che condividono alcune variabili categoriali d'interesse (ad esempio, la nazionalità). Date le possibili disuguaglianze nelle numerosità dei sottoinsiemi, è opportuno condurre le analisi sia sui valori assoluti che su quelli relativi. Per fare ciò, ci si servirà di diverse matrici; l'uso delle matrici permette anche di distinguere il comportamento dei gruppi di nodi come sorgenti o bersaglio di archi, e quindi differenziare produttività e attrattività di ogni *cluster*.

Innanzitutto definiamo la matrice Z, in cui l'elemento z_{ij} è dato dal numero di archi che esce dal gruppo i per terminare nel gruppo j. La matrice Z è quindi composta dalla numerosità assoluta degli archi tra due dati gruppi, e non è affatto robusta rispetto alle numerosità degli stessi.

Dalla matrice Z, si può definire la matrice P, il cui generico elemento è dato da:

$$p_{ij} = \frac{z_{ij}}{|gruppo\ i|} \tag{2.4}$$

dove con $|gruppo\ i|$ si intende la cardinalità, o numerosità, del gruppo i.

In questa maniera, si ottiene la media di connessioni che un membro del gruppo i ha verso il gruppo j.

In modo analogo, si definisce la matrice Q, il cui generico elemento si determina:

$$q_{ij} = \frac{z_{ij}}{|gruppo\ j|} \tag{2.5}$$

Qui la prospettiva è ribaltata: si ricava la media di connessioni ricevute dal gruppo j che provengono dal gruppo i.

Quindi, mentre nel primo caso i valori della matrice indicano la tendenza media di un individuo del gruppo i a connettersi con uno del gruppo j, nel secondo si ottiene quanto il gruppo i contribuisce alle connessioni in entrata del gruppo j. Si noti che i totali di riga di P forniscono il numero medio di archi uscenti dai gruppi

corrispondenti; le righe forniscono quindi la distribuzione media osservata degli archi uscenti da ogni insieme di nodi. Discorso analogo vale per le colonne di Q, che corrispondono alla distribuzione media osservata degli archi entranti in ogni gruppo.

Dal momento che le matrici P e Q restano sensibili alle numerosità dei gruppi, è opportuno definire due ulteriori matrici, i cui generici elementi sono la proporzione di archi rispettivamente uscenti ed entranti nel gruppo:

$$s_{ij} = \frac{z_{ij}}{z_{i+}} \qquad t_{ij} = \frac{z_{ij}}{z_{+j}} \tag{2.6}$$

dove con z_{i+} e z_{+j} si indicano il totale della riga i e della colonna j, rispettivamente.

La prospettiva è analoga a quella già presentata per le matrici P e Q, ma i valori contenuti nelle entrate delle matrici S e T sono proporzioni. Le righe di S e le colonne di T rappresentano dunque le distribuzioni empiriche degli archi, rispettivamente in uscita e in entrata, e pertanto sommano a 1.

L'uso di queste matrici permette di individuare le interazioni più frequenti tra varie porzioni della rete, e di rapportarle sia alle popolosità dei gruppi che alla loro tendenza a creare connessioni.

2.2 Gli indici di rilevanza

2.2.1 Generalità

Quando si tratta una rete, tra gli interessi principali vi è l'identificazione dei nodi più importanti all'interno della stessa. La definizione di "importanza" non è però univoca, e ciò ha causato la proliferazione di indici e misure che provano a tradurre in maniera ottimale questo concetto in linguaggio matematico e statistico. Tali misure, seguendo quanto delineato in Wasserman e Faust (1994), si possono raggruppare nei cosiddetti indici di rilevanza, divisi poi in indici di centralità e indici di prestigio. I primi sono tendenzialmente usati per le reti indirette; i secondi, la cui definizione può essere più complessa e non immediata, sono specifici delle reti dirette. Ciò è dovuto al fatto che in una rete indiretta la tendenza di un nodo a formare archi con altri nodi è equivalente alla sua tendenza ad attrarne; in altre parole, la sua produttività e la sua attrattività coincidono. In una rete diretta i due aspetti sono invece distinti, e possibilmente molto diversi anche sullo stesso nodo.

Per la definizione degli indici di rilevanza ci si basa su due concetti fondamentali: la socialità di un nodo e le distanze tra i nodi della rete. La socialità di un nodo racchiude la sua produttività e la sua attrattività. È intuitivo che un nodo che interagisce molto con altri nodi, e forma quindi molti archi, può assumere un'importanza maggiore nella rete e fungere da ponte tra attori diversi, e magari anche comunità. La distanza (geodesica) tra due nodi n_i e n_j è invece determinata dal numero di archi che si devono percorrere per andare da n_i a n_j. Se il percorso in questione non esiste, si pone convenzionalmente tale distanza infinita; si assume inoltre che due nodi distanti non siano strettamente legati. Poiché spesso i percorsi sono molteplici, si fa spesso riferimento al più breve (*shortest path*). L'uso della distanza permette inoltre di attribuire più rilevanza a nodi non necessariamente molto collegati, ma allo stesso tempo "vicini" a qualunque altro nodo della rete.

Socialità e vicinanza non sono però qualità premiate in tutti gli ambiti: ci sono delle situazioni, ad esempio in economia, in cui il potere contrattuale di un attore è inficiato dal numero di alternative che un suo *partner* ha a disposizione. È quindi possibile che, in determinati contesti, la centralità o il prestigio all'interno di una rete non siano proporzionali al potere che un dato nodo può esercitare. Per alcuni indici che provano a tenere in considerazione questo aspetto, si rimanda a Bonacich (1987).

Spesso, per praticità e comparabilità tra contesti differenti, è uso normalizzare gli indici. In questo elaborato verranno presentati gli indici grezzi, ma si sfrutterà più spesso la versione normalizzata; ad ogni modo, i risultati sono qualitativamente equivalenti.

2.2.2 Alcuni indici di rilevanza

La misura di rilevanza più elementare è il grado (*degree*) di un nodo, ovvero il numero di nodi che gli sono adiacenti. Nel caso di una rete direzionata, si può anche differenziare tra *out-degree*, ovvero il numero di archi che escono dal nodo, e *in-degree*, dato dal numero di archi che invece il nodo riceve. Detti y_{i+} e y_{+i} i totali rispettivamente di riga e colonna della matrice di adiacenza $\mathbf{Y}$, definiamo formalmente grado, *in-degree* e *out-degree* come segue:

$$D(n_i) = y_{i+} + y_{+i}$$
$$ID(n_i) = y_{+i}$$
$$OD(n_i) = y_{i+}$$

Questa misura non tiene però conto del contesto in cui si trovano i nodi: valori elevati di *in-degree* e *out-degree* non implicano che le connessioni in questione siano rilevanti.

La misura più semplice basata sulla distanza è l'eccentricità (*eccentricity*). Essa si definisce come la distanza massima tra il nodo n_i e qualunque altro nodo nella rete; la distanza in questione può tenere conto o meno delle direzioni degli archi. Questo valore è definito solo per reti connesse (fortemente); va quindi valutato con cautela il suo utilizzo. Lo definiamo formalmente:

$$e(n_i) = max\{d(n_i, n_j) : n_j \in \mathcal{N}\}$$

Valori più bassi indicano i nodi in posizioni più strategiche; è evidente come un grado elevato possa abbassare l'eccentricità di un nodo.

Con questa misura possiamo anche definire il raggio di una rete, data dalla sua eccentricità minima, e il diametro, dato dall'eccentricità massima.

Per migliore interpretabilità, ci si serve spesso dell'indice di eccentricità, dato dal reciproco di $e(n_i)$, che dovrebbe risultare proporzionale all'importanza del nodo nella rete:

$$E(n_i) = \frac{1}{e(n_i)}$$

Un indice più complesso, che tiene in considerazione le distanze tra tutti i nodi, è la vicinanza (*closeness*). Si serve dell'inverso delle distanze tra i nodi; valori elevati

di questa misura indicano che un nodo può raggiungere facilmente gli altri all'interno della rete. Formalmente:

$$C(n_i) = \frac{N-1}{\sum_{i \neq j} d(n_i, n_j)}$$

Il grado, che tiene conto solo delle connessioni locali, è interpretativamente affiancato dalla *closeness*, che invece considera la struttura della rete nel suo complesso. Questo indice si può però calcolare solo su componenti connesse. Tale limite è stato superato con l'introduzione della centralità armonica (Rochat, 2009):

$$H(n_i) = \sum_{i \neq j} \frac{1}{d(n_i, n_j)}$$

La formula è analoga a quella della *closeness*, come lo è l'interpretazione, ma l'applicabilità più estesa ne hanno determinato il successo.

Una misura di rilevanza che considera tutti gli *shortest path* all'interno della rete può essere definita intermediazione (*betweenness*). Si definisce tramite il numero di *shortest path* che attraversano il nodo d'interesse. Un attore che è presente in molti percorsi e funge da ponte ha un ruolo importante nel mettere in comunicazione un numero elevato di nodi, e avrà associata una *betweenness* elevata. L'indice è dato dall'espressione:

$$B(n_i) = \sum_{n_j \neq n_i \neq n_k} \frac{\sigma_{n_j n_k}(n_i)}{\sigma_{n_j n_k}}$$

dove $\sigma_{n_j n_k}$ è il numero di *shortest path* tra n_j e n_k e $\sigma_{n_j n_k}(n_i)$ è il numero di *shortest path* tra n_j e n_k che passano per n_i. Si vede dalla definizione che tiene intrinsecamente conto della direzione degli archi. A differenza di *closeness* e centralità armonica, che misurano la vicinanza, la *betweenness* indica quanto un nodo sia necessario per connettere altri nodi. Non viene però tenuto conto dei nodi presenti su percorsi che non siano i più brevi, e nemmeno della "qualità" delle connessioni.

Per provare a tenere conto di quest'ultima, si fa uso della *eigen-centrality*. Questo indice sfrutta il calcolo degli autovalori della matrice di adiacenza per pesare l'importanza di un nodo nella rete, e si presenta come una misura dalle solide basi algebriche:

$$x_i = \frac{1}{\lambda} \sum_j y_{ij} x_j$$

dove x_i è la *eigen-centrality* del nodo n_i, λ è il più grande autovalore della matrice di adiacenza $\mathbf{Y}$, y_{ij} è l'elemento i, j della matrice di adiacenza. Si noti che la somma è, di fatto, solo sui nodi n_j connessi a n_i.

Lo *score* di un nodo è quindi proporzionale allo *score* dei nodi a cui esso è adiacente; ciò porta ad aumentare la rilevanza di nodi in zone dense della rete, talvolta in modo non del tutto coerente con la realtà, se la zona in questione è densa ma poco rilevante. Anche questa misura assume un significato solo in reti connesse. L'affidabilità di questo indice l'ha posto alla base di altre misure di centralità.

Un indice più sofisticato basato sulla *eigen-centrality* è *pagerank*, presentato da Brin e Page (1998) e alla base del motore di ricerca Google. *Pagerank* può essere

applicato su diversi tipi di rete e ha una notevole robustezza. L'algoritmo, che simula il comportamento di un "navigatore casuale", ha dei buoni tempi di convergenza, nonostante lo *score* di un nodo sia dipendente da quello degli altri, come si evince dalla definizione:

$$PR(n_i) = \frac{1-d}{N} + d \cdot \sum_{n_j \in M(n_i)} \frac{PR(n_j)}{OD(n_j)}$$

dove d è il fattore di smorzamento (*damping factor*), tipicamente impostato a 0.85, e $M(n_i)$ è l'insieme dei nodi che hanno un arco verso n_i. È evidente anche come l'accento sia posto sugli archi che entrano in n_i.

Altri indici basati sulla *eigen-centrality* sono quelli calcolati con l'algoritmo Hyperlink-Induced Topic Search (HITS) (Kleinberg, 1999). Per ogni nodo vengono calcolati due indici mutuamente dipendenti, *hub* e *authority*, che rilevano tendenze analoghe alle già presentate produttività e attrattività. Si tratta di misure meno robuste di *pagerank*, ma che permettono di distinguere il valore di un nodo come *hub*, ovvero nella diffusione di archi, o come autorità (*authority*), quindi nell'attrazione degli stessi. La mutua dipendenza fa sì che gli *score* non tengano conto solo della quantità delle connessioni, ma anche della loro qualità:

$$hub(n_i) = \sum_{n_j \in O(n_i)} auth(n_j)$$
$$auth(n_i) = \sum_{n_j \in I(n_i)} hub(n_j)$$

dove $O(n_i)$ è l'insieme dei nodi a cui n_i punta (*out-neighbors*) e $I(n_i)$ è l'insieme dei nodi che puntano a n_i (*in-neighbors*).

2.3 La rappresentazione grafica di una rete

Un problema non banale presentato dalle reti è la loro visualizzazione. In alcuni casi, si opta per la rappresentazione della matrice di adiacenza, ma più frequentemente si raffigura la rete seguendo la convenzione per cui i nodi sono rappresentati da cerchi o altre figure geometriche e gli archi tra di essi sono dati da delle linee, rette o curve. Nel caso di reti dirette, si aggiungono delle frecce alle estremità di tali linee. I nodi possono avere dimensioni differenti, spesso proporzionali a qualche indice (grado o *betweenness*), ma ci si serve anche di forme e colori per evidenziare specifiche variabili d'interesse associate ai nodi. Sebbene ogni rete ammetta infinite rappresentazioni grafiche, solo una piccola frazione di queste può essere considerata realmente efficace per fini esplorativi e interpretativi.

Una rappresentazione grafica efficace riflette fedelmente la struttura sottostante della rete, evidenziando le correlazioni con gli indici di rilevanza e le eventuali comunità. I nodi più importanti tendono a occupare posizioni strategiche nella disposizione spaziale, distinguendosi come *hub* grazie all'elevata concentrazione di archi incidenti. Spesso, le *communities* si manifestano come gruppi di nodi più densamente connessi o spazialmente vicini, e condividono attributi simili, riflettendo fenomeni di omofilia o affinità. L'osservazione di tali corrispondenze può rivelare

importanti dinamiche all'interno della rete, fornendo preziose intuizioni sulla sua struttura e sui processi che la governano. L'uso congiunto di grafici, indici e variabili si rivela quindi un potente strumento esplorativo della rete.

La sfida principale della visualizzazione delle reti è quindi data dalla posizione spaziale assunta dai nodi. Trovare un *set* di coordinate che generi un grafico sia informativo che esteticamente appetibile non è immediato, e ciò ha portato a una gran varietà dei cosiddetti algoritmi di *layout*. Alcuni sono pensati per reti piccole, altri per reti di dimensioni considerevoli, e in molti casi la rappresentazione ottimale dipende dal campo di applicazione. Per valutare la bontà di un *layout* si può considerare la lunghezza totale degli archi disegnati, oppure si conta il numero di volte in cui gli archi si incrociano. La rappresentazione di una rete sotto forma di grafo planare, ovvero privo di intersezioni tra gli archi, è infatti spesso irrealizzabile; tuttavia, gli algoritmi di *layout* mirano ad approssimare questo ideale, minimizzando per quanto possibile gli incroci.

Gli algoritmi in oggetto mirano quindi a facilitare la comprensione e l'analisi delle reti secondo diversi criteri. Esistono varie categorie di algoritmi di *layout*; per un'analisi più approfondita si consiglia di consultare Salter-Townshend *et al.* (2012) e i relativi riferimenti. Alcune di queste tecniche dispongono i nodi seguendo forme geometriche specifiche, come cerchi, stelle o sfere. Altre, pensate per enfatizzare delle strutture gerarchiche, prediligono configurazioni ad albero o affini. Ci sono inoltre approcci che sfruttano la scomposizione spettrale della matrice di adiacenza o tecniche di riduzione dimensionale (*MultiDimensional Scaling*).

Tra gli approcci di *layout*, quelli basati su forze (*force-directed*) hanno guadagnato particolare popolarità. Questi modellano la rete come un sistema dinamico di particelle (i nodi) soggette a forze di attrazione e repulsione (rappresentate dagli archi o dalla loro mancanza). Sebbene non ne siano una replica esatta, queste forze traggono ispirazione da principi fisici: nell'algoritmo Fruchterman–Reingold, ad esempio, i nodi si respingono come cariche elettriche (seguendo la legge di Coulomb), mentre gli archi fungono da molle elastiche tra nodi connessi (in accordo con la legge di Hooke). Il processo inizia da una disposizione casuale e procede iterativamente, alternando fasi di attrazione e repulsione, guidato da parametri predefiniti, detti di *tuning*. Tali parametri aiutano a gestire vari aspetti del processo di *layout*, quali la velocità di convergenza e la sparsità del *set* di coordinate risultante. L'obiettivo è raggiungere uno stato di energia minima, corrispondente alla configurazione più stabile del sistema. I punti di forza di questi algoritmi includono l'abilità di preservare e mettere in risalto simmetrie e strutture comunitarie intrinseche alla rete, unita a una notevole flessibilità e un ampio spettro di applicazione. Tuttavia, presentano anche delle limitazioni: la complessità computazionale è generalmente elevata ($O(n^2)$); inoltre, l'elemento stocastico e la tendenza a convergere verso minimi locali possono portare a risultati variabili e potenzialmente instabili rispetto a piccole alterazioni nella topologia della rete. Spesso si ottengono buoni risultati utilizzando diversi algoritmi in sequenza, in base alle loro qualità.

In questo elaborato verranno impiegati gli algoritmi YifanHu (Hu, 2005) e ForceAtlas2 (Jacomy *et al.*, 2014). Concepiti per l'analisi di reti su larga scala, questi metodi presentano un tempo di convergenza di $O(n \log(n))$, offrendo una notevole efficienza computazionale rispetto ad altri approcci, pur sacrificando in parte la precisione dei risultati.

2.4 Analisi esplorativa della rete Pantheon

2.4.1 Descrizione della rete globale

Per le seguenti analisi ci si è serviti, in `R`, del pacchetto `igraph` (Csardi e Nepusz (2006), Csárdi *et al.* (2024)).

La rete Pantheon si distingue per la sua considerevole dimensione, sia in termini di nodi che di archi. Un aspetto notevole è che la quasi totalità dei nodi appartiene a una componente debolmente connessa (Tabella 2.1). I pochi attori che ne rimangono esclusi sono caratterizzati da un grado basso o nullo, suggerendo una rilevanza trascurabile nel contesto globale della rete.

Nonostante il grado medio dei nodi sia considerevole, la densità della rete (equazione 2.1) risulta essere minima. Si parla dunque di rete sparsa: il numero di archi osservati è relativamente modesto rispetto al numero di nodi, suggerendo connessioni selettive piuttosto che una struttura densamente interconnessa. Il valore del grado medio sembra inoltre influenzato da pochi nodi molto connessi. La mediana, indicatore più robusto, è infatti sensibilmente inferiore alla media (Figura 2.1).

La presenza di una certa reciprocità (equazione 2.2) indica che una frazione non trascurabile delle interazioni tra gli attori tende a essere bidirezionale. La transitività indica una certa tendenza alla formazione di *cluster* o "triangoli" di connessioni.

Questi risultati sembrano coerenti con le dinamiche alla base della rete. Tuttavia, è importante considerare l'impatto del *selection bias*. La soglia arbitraria delle 25 lingue potrebbe aver escluso personaggi cruciali per la topologia locale del sistema, potenzialmente riducendo la transitività osservata. Questa ipotesi trova supporto nel diametro piuttosto elevato della rete, che suggerisce percorsi talvolta tortuosi tra i nodi.

La rete Pantheon si presenta come un sistema sfaccettato, in cui la sparsità globale coesiste con aree di maggiore densità locale. È quindi logico considerare anche alcune porzioni della rete nelle analisi, poiché le caratteristiche topologiche potrebbero mostrare una variabilità notevole.

2.4.2 Analisi per sottoreti

Si è già evidenziato come i nodi della rete presentino eterogeneità, sia nella loro collocazione spazio-temporale che negli ambiti occupativi (Capitolo 1). Le porzioni di rete costituite da nodi accomunati dal continente, Paese o secolo di nascita, o appartenenti al medesimo dominio, presentano infatti coefficienti di densità, reciprocità e transitività superiori rispetto a quelli della rete considerata nel suo complesso. Questa considerazione rafforza le precedenti osservazioni che suggeriscono omofilia e strutture di comunità.

Più in particolare, le reti di personalità provenienti dallo stesso continente presentano una densità inversamente proporzionale al numero di nodi (Figura 2.2), in linea con la struttura della formula (2.1). La reciprocità si rivela particolarmente elevata tra figure africane, asiatiche e, soprattutto, dalla provenienza sconosciuta. Quest'ultimo gruppo presenta anche il livello di transitività di gran lunga più alto, mentre nordamericani ed europei mostrano questa caratteristica in maniera decisamente sotto la media. È plausibile che anche la gran quantità di unità statistiche

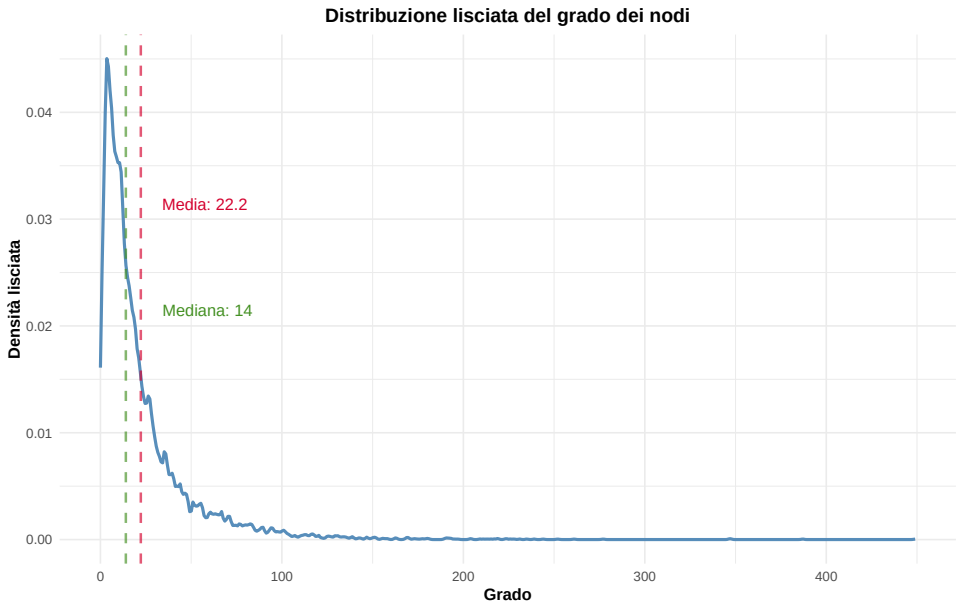


Figura 2.1: Distribuzione empirica lisciata del grado dei nodi della rete Pantheon

Nodi	**Nodi connessi**	**Archi**	**Diametro**	**Percorso medio**
11340	11062	126153	15	4.81
Densità	**Reciprocità**	**Transitività**	**Grado medio**	**Grado mediano**
0.000981	0.346	0.156	22.2	14

Tabella 2.1: Statistiche rilevate sulla rete

provenienti da questi continenti, e la conseguente sparsità, abbiano influenzato queste misure.

Approfondendo l'analisi sulle singole nazioni, continua a valere la stessa tendenza evidenziata in precedenza per la densità delle sottoreti: i Paesi cui corrispondono meno unità statistiche sono quelli più densamente collegati (Figura 2.3). La reciprocità è particolarmente alta tra individui di nazionalità ceca, canadese e indiana; la probabilità che una connessione venga ricambiata supera il 50%. Gli individui di provenienza indiana mostrano anche la più alta transitività. Gli Stati Uniti, invece, presentano coefficienti piuttosto modesti per entrambe le statistiche: anche in questo caso, la numerosità di nodi sembra favorire la sparsità. È inoltre curiosamente bassa la quantità di connessioni reciproche tra individui brasiliani.

La rilevazione di densità, reciprocità e transitività sulle sottoreti date dai secoli di nascita offre risultati analoghi (Figura 2.4): i secoli più remoti sono caratterizzati da meno unità statistiche e meno sparsità, e conseguentemente da coefficienti più elevati. Il comportamento di questi indici nel tempo ha un andamento abbastanza omogeneo, eccezion fatta per i secoli XI e XIII, che mostrano grandi transitività e reciprocità, rispettivamente.

Nell'analisi degli individui accomunati dallo stesso dominio, la densità conferma nuovamente la sua dipendenza dal numero di nodi della sottorete (Figura 2.5). Più interessanti sono invece i risultati su reciprocità e transitività: le figure umanistiche e artistiche mostrano coefficienti relativamente bassi, che suggeriscono una ridotta influenza nelle interazioni tra nodi. Al contrario, gli sportivi mostrano valori di

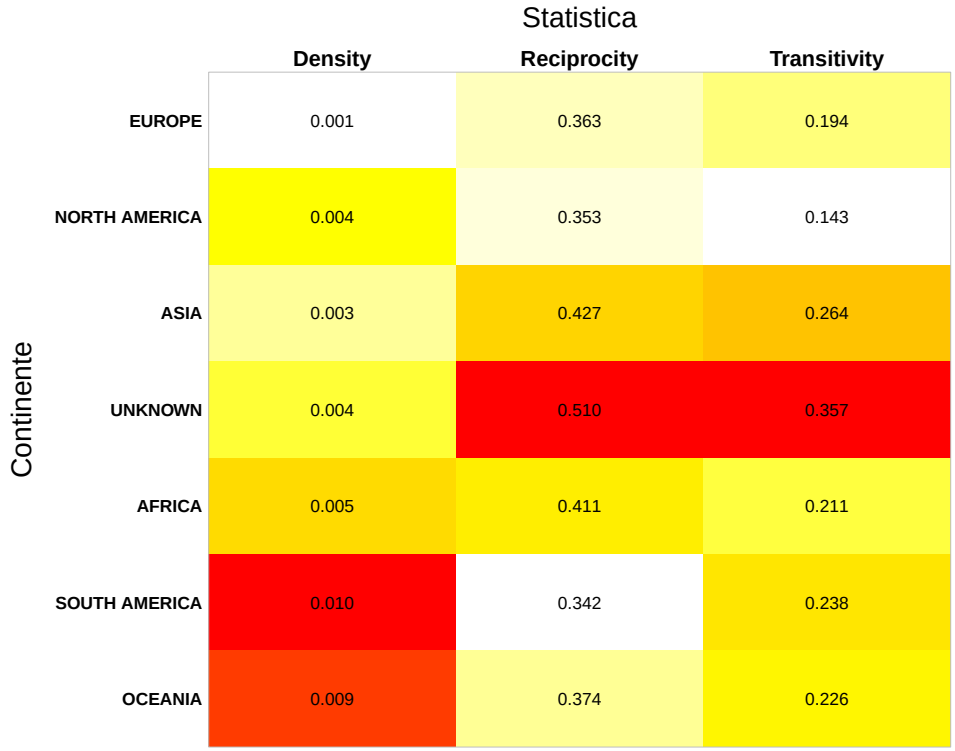


Figura 2.2: Densità, reciprocità e transitività nei continenti

reciprocità e transitività tra i più elevati, nonostante la considerevole popolosità del dominio suggerisse diversamente. Questo contrasto evidenzia differenze sostanziali nella struttura delle relazioni all'interno dei domini, che sembrano meno legate alle popolosità e più a una natura intrinseca degli stessi.

Infine, anche prendendo in esame le sottoreti formate solo da uomini e solo da donne si ripresenta la stessa dipendenza tra le statistiche e l'ordine della sottorete. Le figure femminili, infatti, costituiscono una rete più densa e interconnessa rispetto agli uomini (Tabella 2.2). Non sembra però possibile trarre conclusioni su eventuali differenze di genere, a causa del grande squilibrio nelle numerosità.

L'analisi delle sottoreti rivela anche la diversa propensione di alcuni gruppi di nodi a ricevere o creare connessioni. Sebbene in media queste due tendenze sembrino essere equilibrate, ci sono delle eccezioni.

I personaggi nordamericani ed europei sono quelli con il grado mediamente più elevato, sia in entrata che in uscita, indicando una maggiore connettività (Figura 2.6). Al contrario, gli individui dalla provenienza sconosciuta mostrano una socialità particolarmente bassa; gli altri quattro continenti esibiscono valori simili tra loro sia per l'*in-degree* che per l'*out-degree*. In questo caso, quindi, sembra suggerita una possibile correlazione tra l'ordine della sottorete e il grado medio interno alla stessa.

Tuttavia, un'analisi più approfondita basata sui Paesi di nascita smentisce questa apparente correlazione (Figura 2.7). Austria, Grecia e Canada sono infatti nazioni i cui individui mostrano una notevole socialità, nonostante non siano tra i Paesi più rappresentati nel *dataset*. In particolare, austriaci e greci tendono ad attrarre più connessioni di quante ne creino, mentre i canadesi appaiono più propensi a stabilire connessioni con altri nodi. All'estremità opposta dello spettro, i personaggi di origine

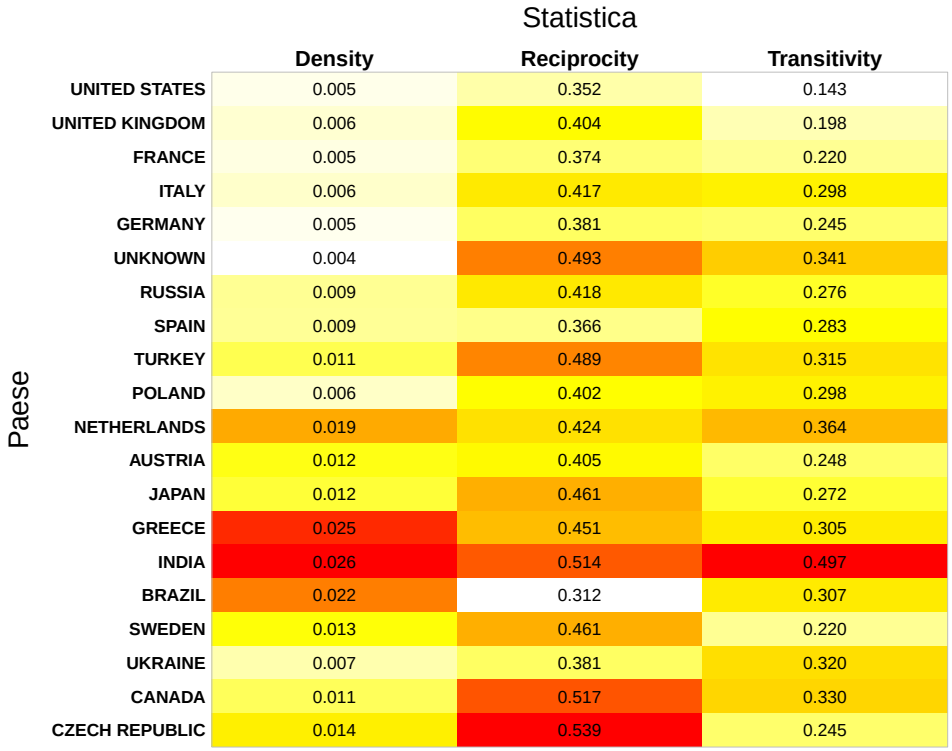


Figura 2.3: Densità, reciprocità e transitività nei Paesi

	Maschi	**Femmine**
Density	0.001	0.005
Reciprocity	0.345	0.420
Transitivity	0.166	0.420

Tabella 2.2: Densità, reciprocità e transitività nei generi

giapponese o sconosciuta sono di gran lunga quelli che presentano meno connessioni, in ambo le direzioni.

L'andamento del grado medio al variare del secolo di nascita presenta un andamento peculiare (Figura 2.8). Il I secolo d.C. emerge come il periodo con il grado medio più elevato; l'andamento di questa statistica segue poi un declino progressivo fino al VII secolo. Successivamente, si osserva una ripresa che culmina nel XX secolo, il secondo più interconnesso; quest'osservazione vale sia per l'*out-degree* che per l'*in-degree*.

L'alto grado medio dei secoli più remoti può essere attribuito alla presenza di un numero limitato di individui, principalmente figure cristiane, che hanno mantenuto la loro rilevanza attraverso i secoli. D'altra parte, la maggiore presenza di individui di epoca moderna nel *dataset* favorisce un aumento delle connessioni tra contemporanei, innalzando il grado medio nei secoli più recenti.

È però, nuovamente, suddividendo i nodi nei rispettivi domini che si ottengono i risultati più interessanti (Figura 2.9). In questo caso, infatti, non sembra esserci alcun legame tra le popolosità dei settori e la tendenza dei nodi a connettersi. Personaggi dei mondi artistico, sportivo e umanistico sono quelli che tendono a interagire di più, al contrario di esploratori e scienziati, il cui grado medio di connessioni è meno della

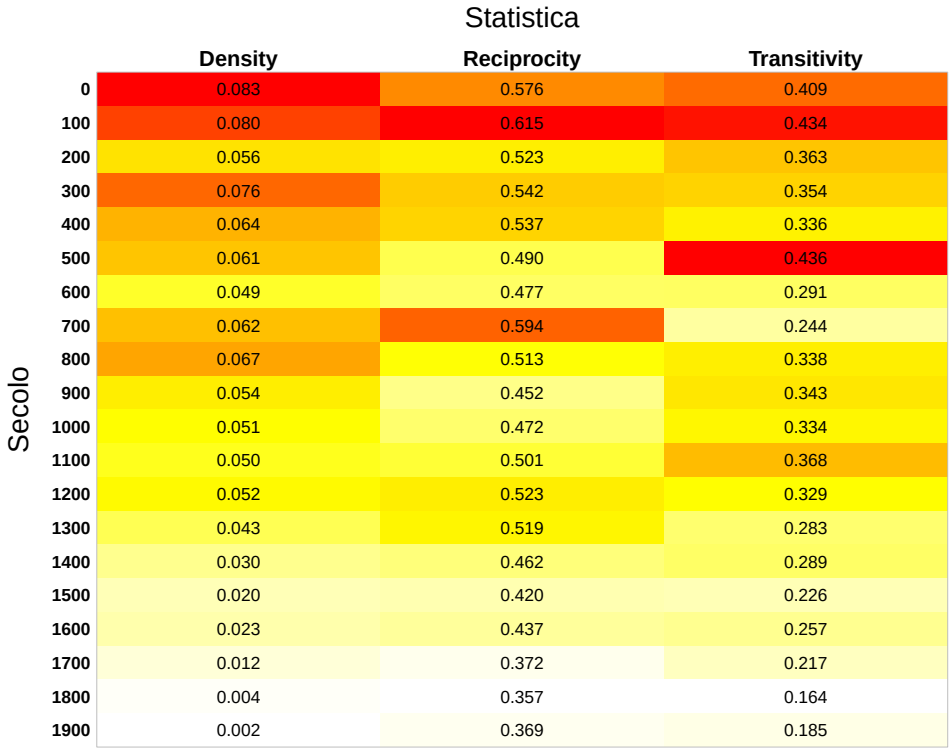


Figura 2.4: Densità, reciprocità e transitività nei secoli

metà rispetto ai gruppi più connessi. Più che un *pattern* universale, sembra emergere una tendenza specifica dei singoli domini, apparentemente slegata dalle numerosità. Va notato, tuttavia, che arte e sport sono domini composti prevalentemente da individui del secolo scorso, e che quindi la contemporaneità tra soggetti, unita alla prossimità cronologica ai nostri tempi, potrebbe aver favorito la connettività di questi nodi.

Le donne hanno un grado medio decisamente superiore a quello degli uomini (Tabella 2.3). Ciò è dovuto non tanto all'*in-degree*, che risulta sostanzialmente uguale, quanto all'*out-degree*. Sembra infatti che le figure femminili tendano a creare molti più legami degli uomini. Non c'è però modo di dire quale dei due generi tenda maggiormente a essere un *authority*.

Dopo aver osservato il comportamento delle varie sezioni della rete, è d'interesse comprendere anche come queste interagiscano. Per farlo, ci si servirà della matrice Z definita nella Sezione 2.1.4, e delle relative trasformazioni ivi presentate (equazioni 2.4, 2.5, 2.6). La visualizzazione delle matrici P, Q e T appesantirebbe la lettura ed è dunque riportata in Appendice A.

L'elevata numerosità di europei e nordamericani si riflette anche nel grande numero di archi di cui essi sono sia sorgente, sia bersaglio (Figure 2.10 e 2.11). L'Europa riceve mediamente una porzione cospicua di archi uscenti da ogni continente, risultando fortemente attrattiva. I continenti meno numerosi sono anche quelli che ricevono tendenzialmente meno archi. Inoltre, le interazioni tra personalità nordamericane, sudamericane e oceaniane con individui dalla provenienza ignota sono pressoché nulle: ciò è probabilmente dovuto alla poca contemporaneità cronologica di questi soggetti.

Nello studio della matrice S indotta dalla suddivisione dei nodi nei continenti

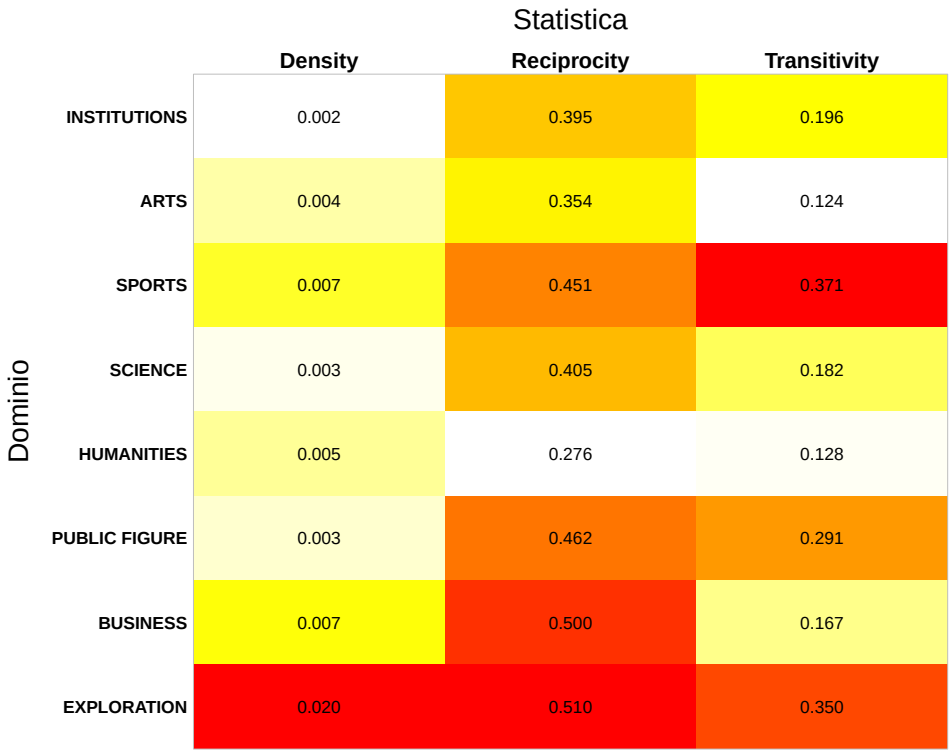


Figura 2.5: Densità, reciprocità e transitività nei domini

	Degree	**In-Degree**	**Out-Degree**
Maschi	21.7	11.1	10.6
Femmine	25.6	11.0	14.6

Tabella 2.3: *Degree*, *in-degree* e *out-degree* medi nei generi

natii, merita particolare attenzione la diagonale: essa indica infatti la proporzione di archi uscenti ed entranti nello stesso continente. La presenza di valori superiori alla media sulla diagonale evidenzia l'autoreferenzialità presente nei vari continenti. Europei e nordamericani sono di gran lunga le figure più autoreferenziali: circa tre quarti delle interazioni di entrambi i continenti sono autoriferite. In generale, comunque, questo fenomeno sembra diffuso, coerentemente con l'ipotesi di omofilia già avanzata in precedenza.

Un'analisi estesa alle nazionalità mostra un comportamento analogo: nella maggior parte delle interazioni, in entrambe le direzioni, sono coinvolti individui provenienti dai 5 Paesi più presenti, ovvero Stati Uniti, Regno Unito, Francia, Italia e Germania (Figure 2.12 e 2.13). Le personalità di queste nazioni sono anche le principali destinatarie di archi, in maniera più o meno indipendente dalla nazionalità del nodo sorgente. Risalta inoltre un'affinità sopra la media tra individui originari di Paesi storicamente legati, quali Stati Uniti e Canada, Germania e Polonia o Russia e Ucraina.

Anche qui, l'affinità tra connazionali è marcata, in particolare nei Paesi dalla numerosità maggiore nel *dataset*. Tuttavia, questo fenomeno sembra attenuato nei territori che non hanno sempre goduto di indipendenza nel corso della storia, quali le già menzionate Ucraina e Polonia.

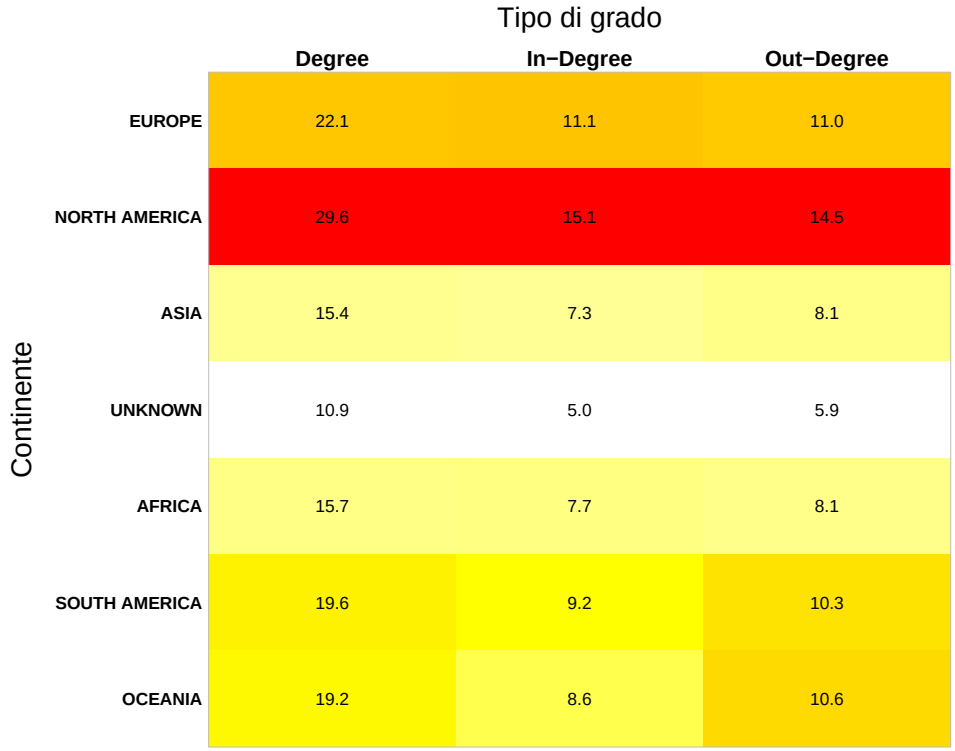


Figura 2.6: *Degree*, *in-degree* e *out-degree* medi nei continenti

Dato l'aumento pseudo-esponenziale di unità statistiche nel corso dei secoli, una trasformazione logaritmica facilita l'analisi delle interazioni tra gruppi definiti dai secoli di nascita (Figura 2.14). In questo caso, si nota come non vi sia solo elevata affinità tra individui nati nello stesso secolo, ma anche con persone nate in secoli attigui. Questo fenomeno si rivela in valori grandi non solo sulla diagonale delle matrici Z e S, ma anche nelle celle a essa adiacenti. Inoltre, si conferma la tendenza dei gruppi più numerosi ad attirare più interazioni (Figura 2.15): sono infatti i secoli più recenti a ricevere una maggiore proporzione di archi in entrata. I nati in tali secoli sono inoltre i più autoaffini, mediamente.

Analizzando i *link* tra i personaggi classificati nei rispettivi domini, le figure istituzionali e artistiche emergono come le più coinvolte e attrattive nelle interazioni (Figure 2.16 e 2.17). Sorprendentemente, le figure umanistiche sembrano attirare molte interazioni, nonostante la loro numerosità non sia tra le più alte. In generale, si osserva anche qui un'elevata autoreferenzialità, accentuata nei domini più popolosi ma soprattutto nell'ambito sportivo. Gli sportivi, difatti, presentano una connettività estremamente bassa con altri domini, suggerendo una tendenza all'isolamento. La grande omofilia in ambito sportivo e artistico va di pari passo con quella osservata all'interno del XX secolo, data la prevalenza di tali figure in questo periodo storico.

Infine, la grande differenza nel numero di uomini e donne si riflette nelle interazioni totali rilevate (Tabella 2.4). Si ricorda, infatti, che a ogni donna corrispondono oltre 6 uomini. Le donne si collegano agli uomini più di quanto non succeda il contrario: le figure maschili, infatti, interagiscono decisamente poco con le proprie controparti femminili (Tabella 2.4). Entrambi i generi presentano autoaffinità. È vero che in apparenza sembra esserci equilibrio nelle distribuzioni degli archi che coinvolgono

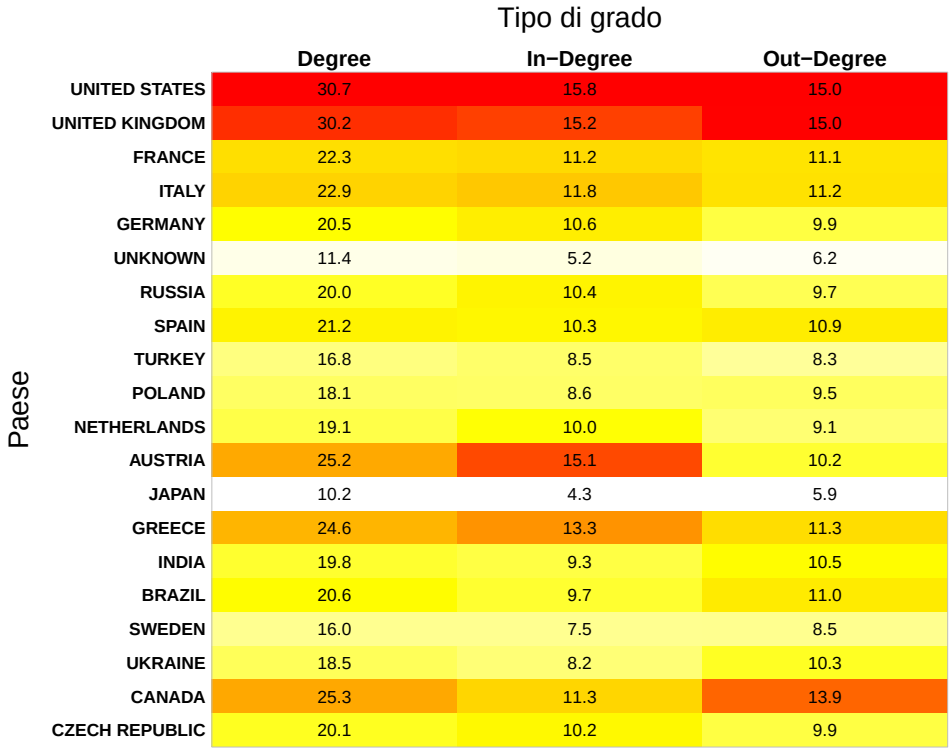


Figura 2.7: *Degree*, *in-degree* e *out-degree* medi nei Paesi

donne; tale ipotesi è però smentita se si considera il già menzionato divario nelle numerosità.

Le osservazioni sull'apparente omofilia caratterizzante la rete sono confermate dai valori assunti dall'assortatività (equazione 2.3). Le sezioni della rete sono state definite secondo le variabili categoriali utilizzate in precedenza, ovvero continente, Paese, secolo di nascita e dominio d'appartenenza (Tabella 2.5). In particolare, i nodi della rete Pantheon mostrano maggiore affinità verso gli individui con cui condividono l'ambito occupazionale e il secolo di nascita. La provenienza geografica, sebbene meno determinante, mantiene comunque un'influenza non trascurabile sulle connessioni. Lo stesso vale per il genere: non si può certo affermare che non condizioni le interazioni, ma tra le variabili analizzate sembra essere la meno influente.

2.4.3 I nodi più rilevanti

La scelta di un indice di rilevanza per individuare i nodi più importanti all'interno di una rete non è sempre semplice. Alle diverse misure corrispondono infatti diverse basi teoriche, e quindi diverse interpretazioni. Non si cerca il "migliore" indice di rilevanza, ma quello "ottimale"; tale aggettivo assume un significato diverso a seconda dello specifico contesto in cui viene applicato. È anche per questo che spesso è incoraggiato l'uso congiunto di molteplici indici. Tale utilizzo è agevolato dall'uso di grafici, che aiutano anche a evidenziare quanto siano concordi le misure stesse. I risultati possono essere molto differenti, come evidenziato in Tabella 2.6.

La proliferazione di indici disponibili ha motivato Beytía e Schobin (2018) a introdurre il BCI, già menzionato sommariamente nella Sezione 1.1.2. Questa misura

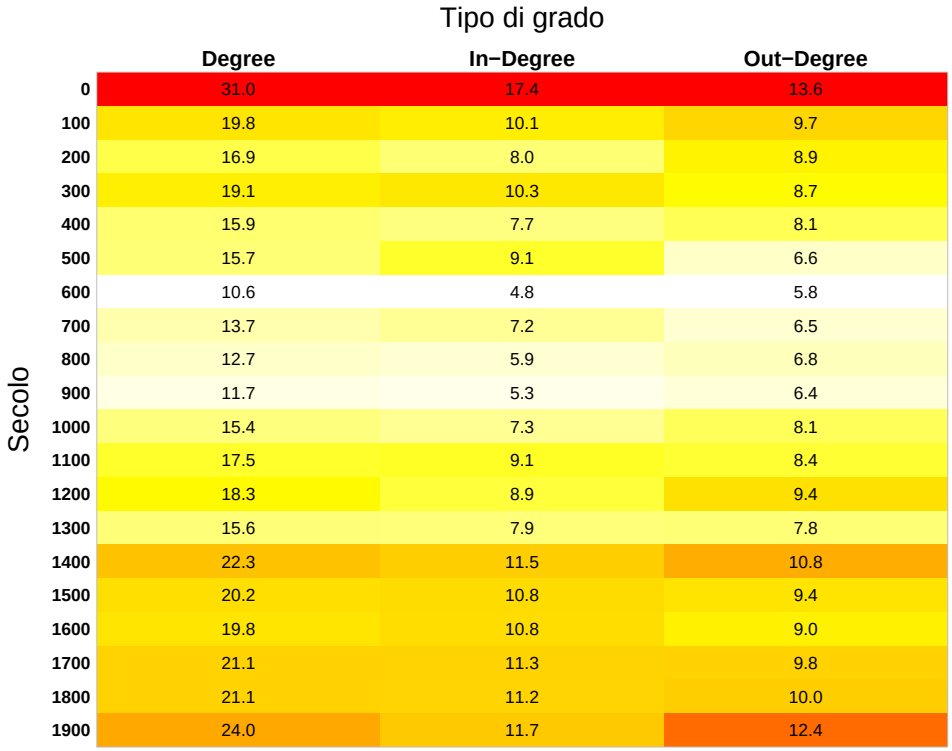


Figura 2.8: *Degree*, *in-degree* e *out-degree* medi nei secoli

	M	F
M	96114	8201
F	13535	8303

	M	F
M	0.92	0.08
F	0.62	0.38

	M	F
M	0.88	0.50
F	0.12	0.50

Tabella 2.4: da sinistra a destra, matrice Z, S, T per genere

combina il *pagerank* di un dato nodo per il numero di lingue in cui la sua pagina Wikipedia è disponibile, in modo da premiare i nodi più interculturali. L'indice viene poi normalizzato, come è consuetudine. Formalmente, detto $NL(n_i)$ il numero di lingue in cui la biografia del nodo n_i è disponibile:

$$BCI(n_i) = \frac{NL(n_i) \cdot PR(n_i) - \min_j \{NL(n_j) \cdot PR(n_j)\}}{\max_j \{NL(n_j) \cdot PR(n_j)\} - \min_j \{NL(n_j) \cdot PR(n_j)\}}$$

Nell'analisi della rete Pantheon, si osserva una certa concordanza tra grado e *betweenness* per i nodi più rilevanti (Figura 2.18). Personaggi come Barack Obama, Adolf Hitler o George Bush sono caratterizzati sia da un grado elevato che da una *betweenness* relativamente alta. Altri individui, come Papa Giovanni Paolo II, David Beckham e Carlo Magno, risultano fondamentali per molti percorsi interni alla rete, più di quanto il loro grado suggerirebbe. In generale, Pantheon si compone di una gran quantità di nodi che presentano sia grado che *betweenness* modesti, a fronte di pochi attori che esercitano un ruolo dominante nelle dinamiche della rete.

Sembrano essere ancora più concordi grado e *pagerank*: i personaggi più connessi risultano tra i più rilevanti anche per il complesso algoritmo. Ciò indica che, mediamente, le interazioni degli attori più attivi non sono solo numerose, ma anche

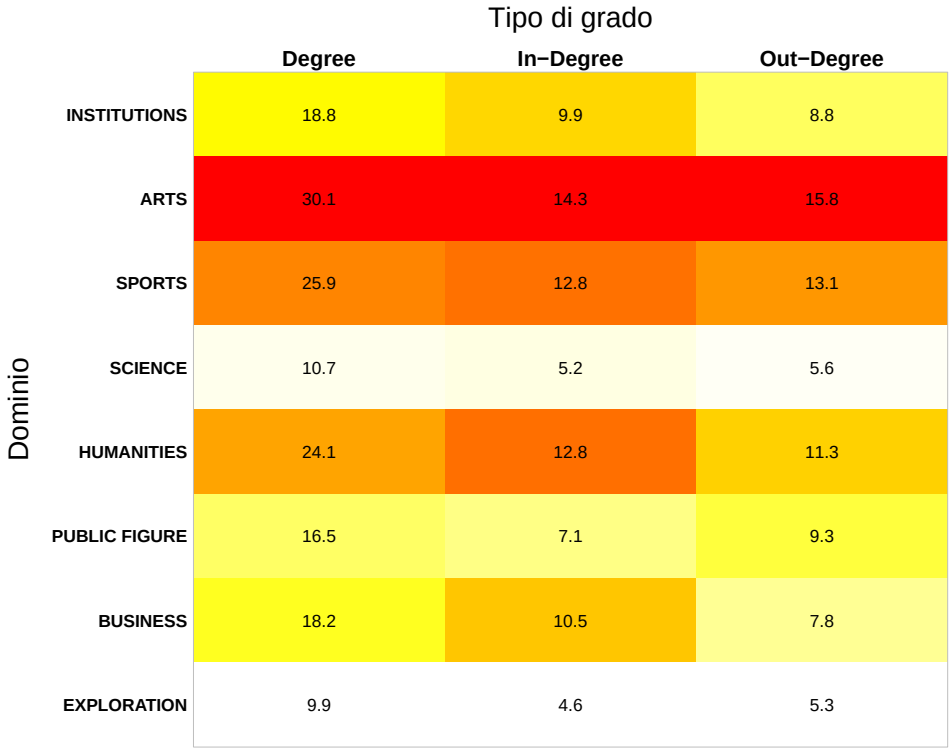


Figura 2.9: *Degree*, *in-degree* e *out-degree* medi nei domini

Raggruppamento	Continente	Paese	Secolo	Dominio	Genere
Assortatività	0.48	0.33	0.61	0.69	0.334

Tabella 2.5: Assortatività calcolata secondo diversi raggruppamenti

qualitativamente rilevanti. Inoltre, a parità di grado, le figure istituzionali appaiono mediamente più importanti di quelle appartenenti ad altri domini.

Il confronto di *pagerank* e *betweenness* rivela che la centralità nella topologia della rete non è sempre associata ad altrettanta importanza secondo l'algoritmo ideato da Brin e Page. Inoltre, sebbene asimmetrica, la distribuzione di *pagerank* appare più omogenea rispetto a quella della *betweenness*. Quest'ultima presenta una concentrazione di valori elevati in un numero minore di nodi, risultando in una distribuzione con una coda destra più lunga.

È invece singolare ciò che rivela un confronto tra *pagerank* e *eigen-centrality*. Mentre il primo indice evidenzia l'importanza di personaggi storici che sono indubbiamente rilevanti, la seconda misura presenta un evidente *bias* che favorisce tennisti uomini del circuito ATP. È possibile che questa distorsione sia dovuta a una grande densità di questa zona della rete: il tennis, infatti, è uno sport individuale dove i giocatori del circuito si affrontano un gran numero di volte in carriera. Il numero elevato di connessioni può dunque generare una sorta di *feedback loop* positivo, per cui l'importanza dei nodi si influenza reciprocamente. È curioso inoltre come i risultati della *eigen-centrality* siano più concordi con *pagerank*, se non si considerano i tennisti. Ad ogni modo, non sembra indicato l'uso dell'indice basato sugli autovalori in questo specifico caso.

Emerge quindi un legame tra gli indici. La correlazione lineare (di Pearson, o

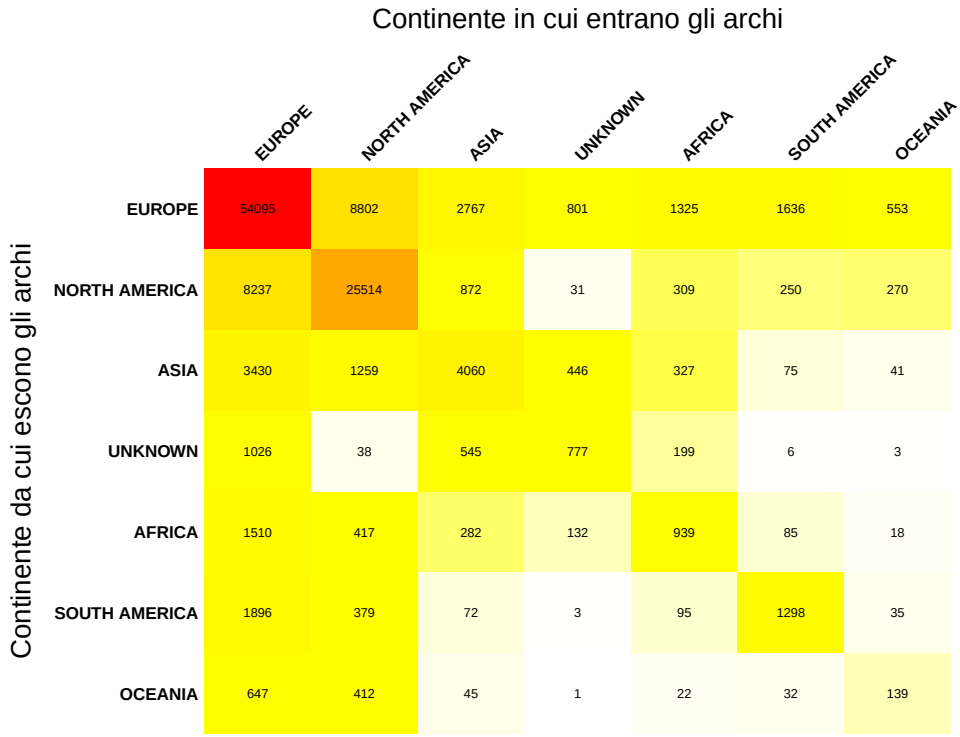


Figura 2.10: Numero di archi in base al continente

ρ) è particolarmente elevata tra alcune misure: il grado è molto legato allo *score pagerank*, oltre che ovviamente ad *out-degree* e, maggiormente, *in-degree*. Anche la *betweenness* è legata agli indici appena menzionati. L'algoritmo HITS è basato sulla scomposizione spettrale della matrice di adiacenza: le comuni basi teoriche si riflettono in un'elevatissima correlazione tra gli *score hub*, *authority* e di *eigen-centrality*. La centralità armonica non mostra legami forti con nessun altro indice, eccezion fatta per l'*out-degree*. Intuitivamente, un nodo con più percorsi in uscita impiega meno "passi" a raggiungere gli altri nodi, risultando in una distanza media più bassa e quindi una centralità armonica più elevata.

L'uso del coefficiente di correlazione lineare è però poco robusto rispetto alle distribuzioni degli indici; queste, infatti, sono molto eterogenee, nonostante tendano a essere tutte asimmetriche a destra. Una misura di correlazione più indicata per questa situazione può essere il coefficiente tau di Kendall (τ), pensato per misurare la concordanza tra coppie di valori.

La correlazione di Kendall riporta stime più elevate dove ρ è prossimo allo zero; in maniera speculare, τ riporta valori mediamente più bassi dove la correlazione lineare si avvicina a uno. In generale, le coppie di indici più correlate non cambiano in maniera rilevante. Va però osservato che la correlazione di Kendall è più robusta rispetto alla rilevanza dei tennisti per l'*eigen-centrality*, e sembra rilevare una certa concordanza generale tra le misure (Figura 2.19).

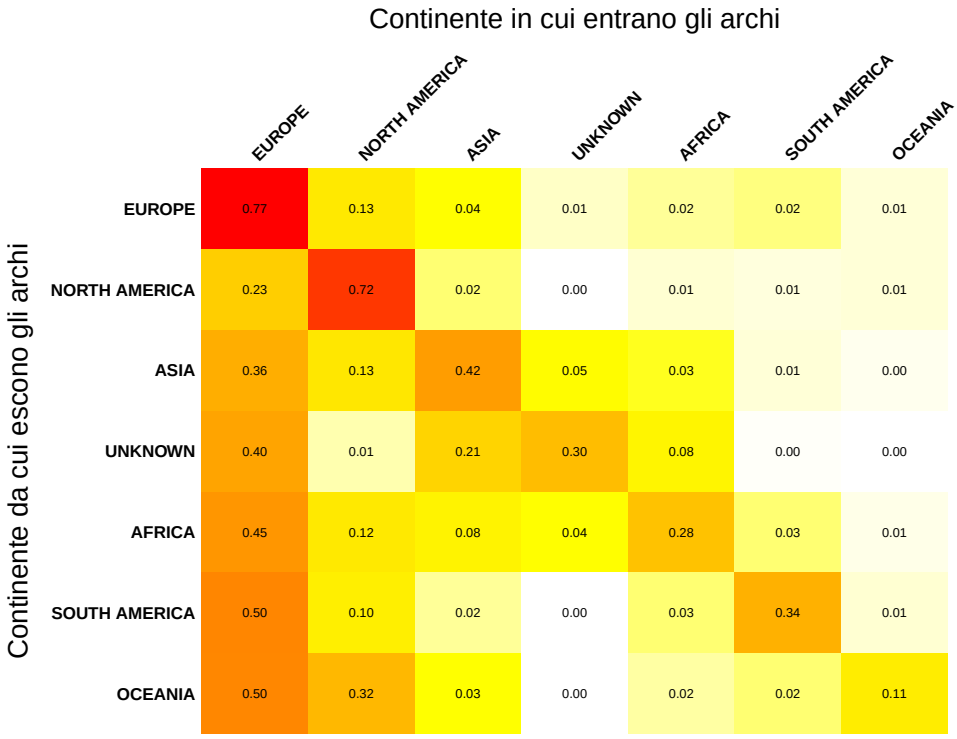


Figura 2.11: Numero di archi in base al continente, relativamente al numero di archi del continente sorgente

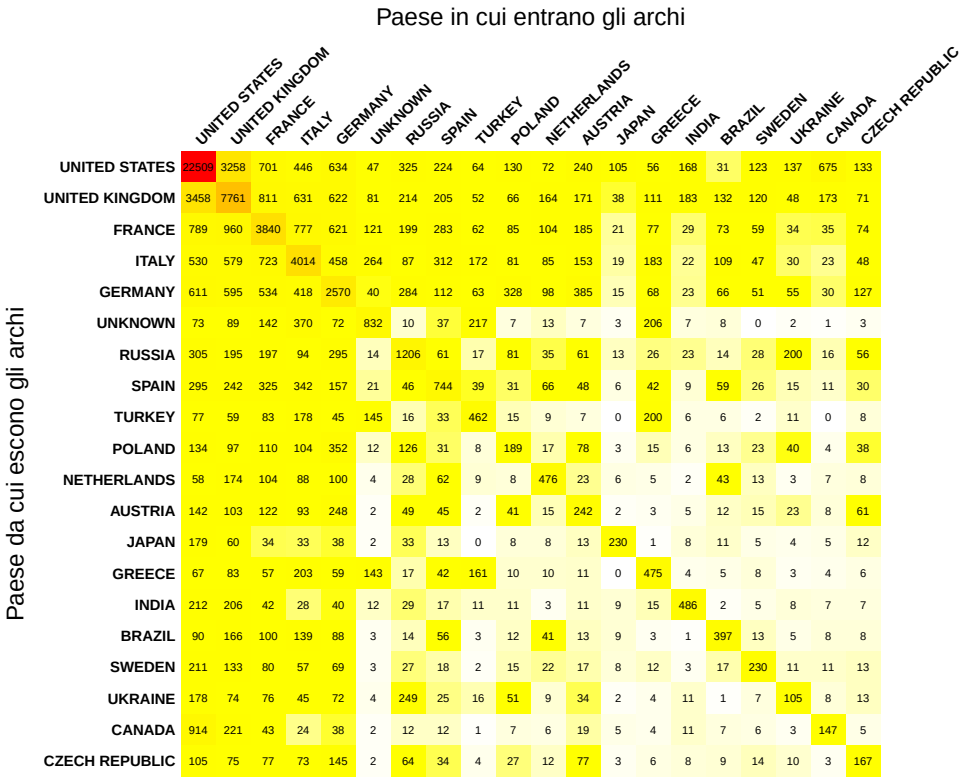


Figura 2.12: Numero di archi in base al Paese

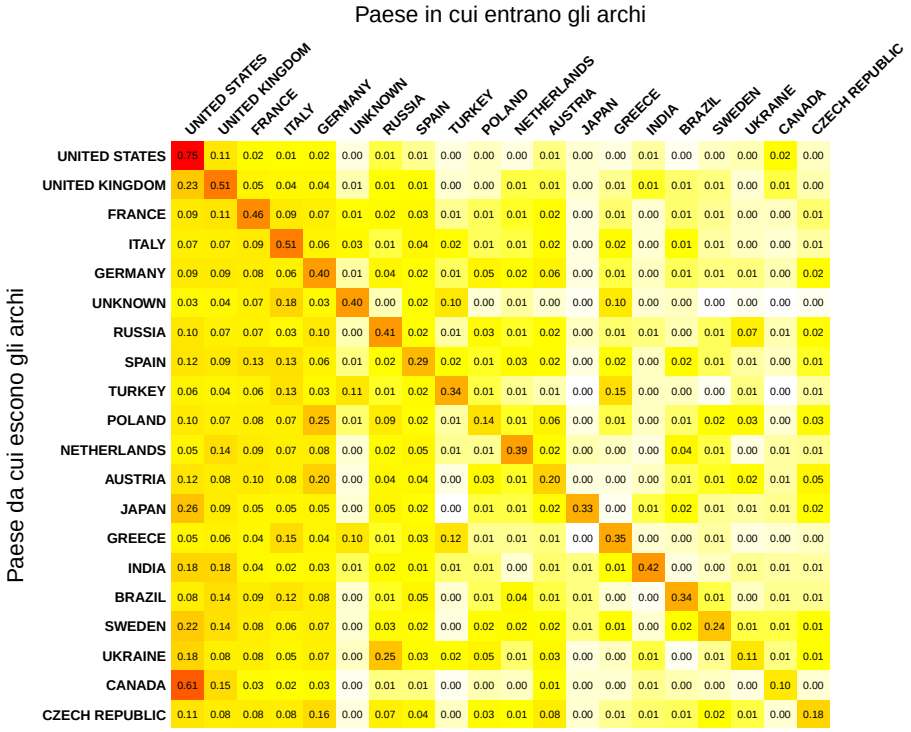


Figura 2.13: Numero di archi in base al Paese, relativamente al numero di archi del Paese sorgente

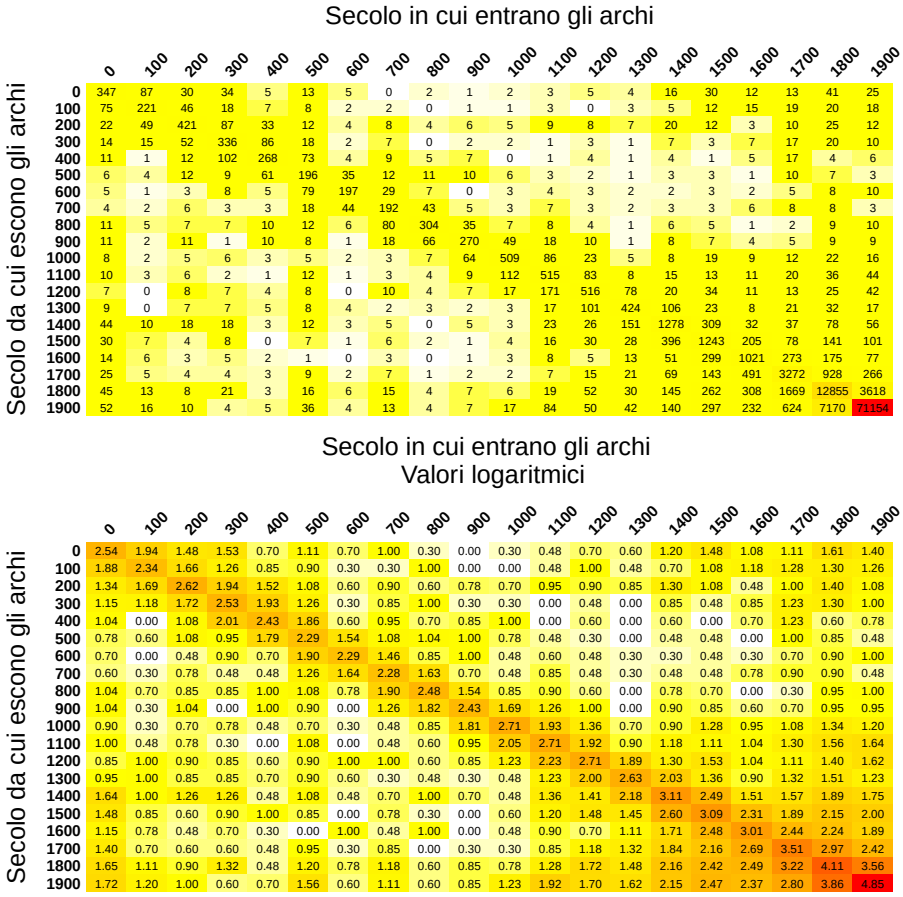


Figura 2.14: Numero e log-numero di archi in base al secolo

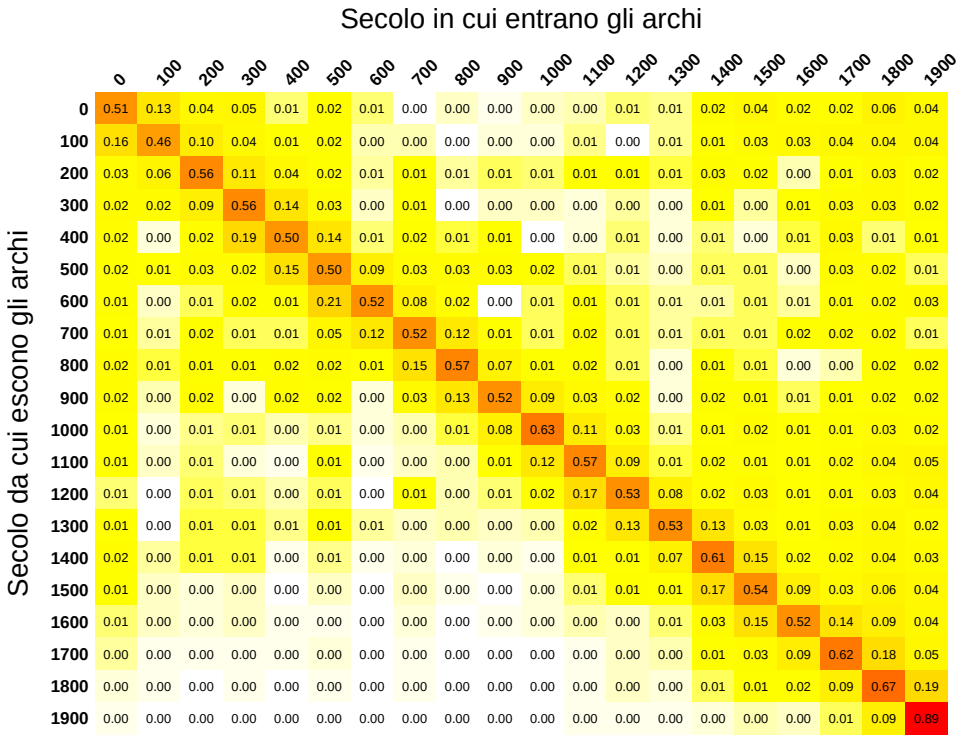


Figura 2.15: Numero di archi in base al secolo, relativamente al numero di archi del secolo sorgente

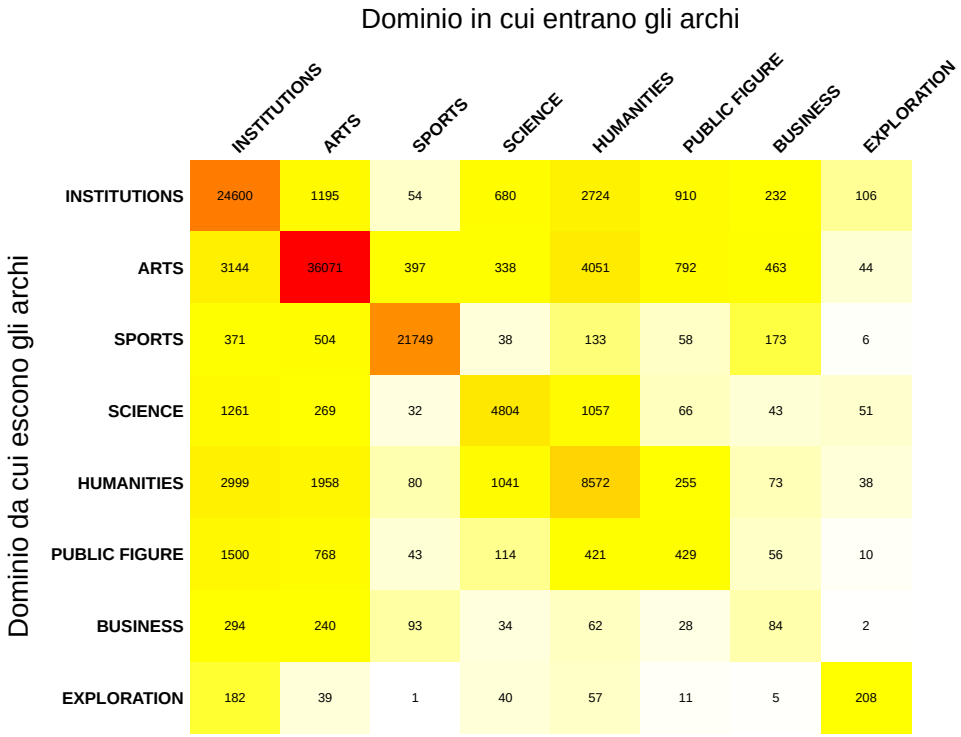


Figura 2.16: Numero di archi in base al dominio

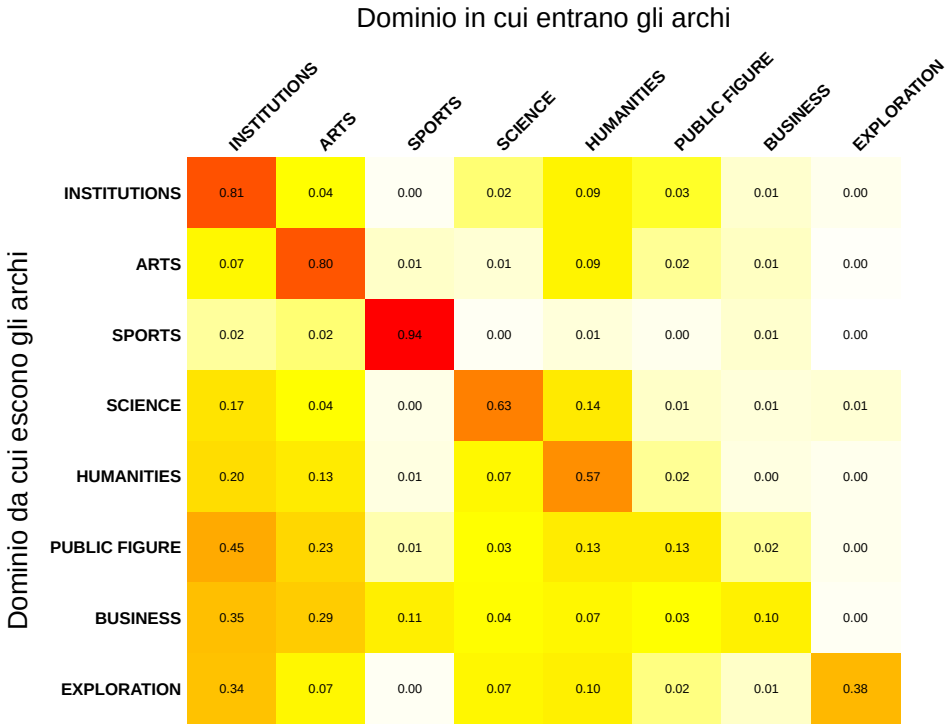


Figura 2.17: Numero di archi in base al dominio, relativamente al numero di archi del dominio sorgente

Rank	**Degree**	**Betweenness**	**PageRank**
1	Barack Obama	Pope John Paul II	Adolf Hitler
2	Adolf Hitler	David Beckham	Barack Obama
3	William Shakespeare	Adolf Hitler	George Bush
4	George Bush	Charlemagne	William Shakespeare
5	Joseph Stalin	George Bush	Joseph Stalin
6	Ronald Reagan	Benito Mussolini	Ronald Reagan
7	Bob Dylan	Albert Einstein	Plato
8	Martin Scorsese	Vladimir Putin	Bill Clinton
9	Bill Clinton	Pelé	Winston Churchill
10	John F. Kennedy	Ronald Reagan	Aristotle
Rank	**Harmonic**	**Eigen**	**HPI**
1	Bob Dylan	Roger Federer	Aristotle
2	Martin Scorsese	Rafael Nadal	Plato
3	Meryl Streep	Andy Roddick	Jesus Christ
4	Ennio Morricone	Novak Djokovic	Socrates
5	Stanley Kubrick	Andrew Murray	Alexander the Great
6	Christopher Nolan	Tomáš Berdych	Leonardo da Vinci
7	Gary Oldman	Jo-Wilfried Tsonga	Confucius
8	Philip Glass	James Blake	Julius Caesar
9	Steven Spielberg	David Ferrer	Homer
10	Benedict Cumberbatch	Andre Agassi	Pythagoras

Tabella 2.6: Nodi principali secondo diversi indici di rilevanza

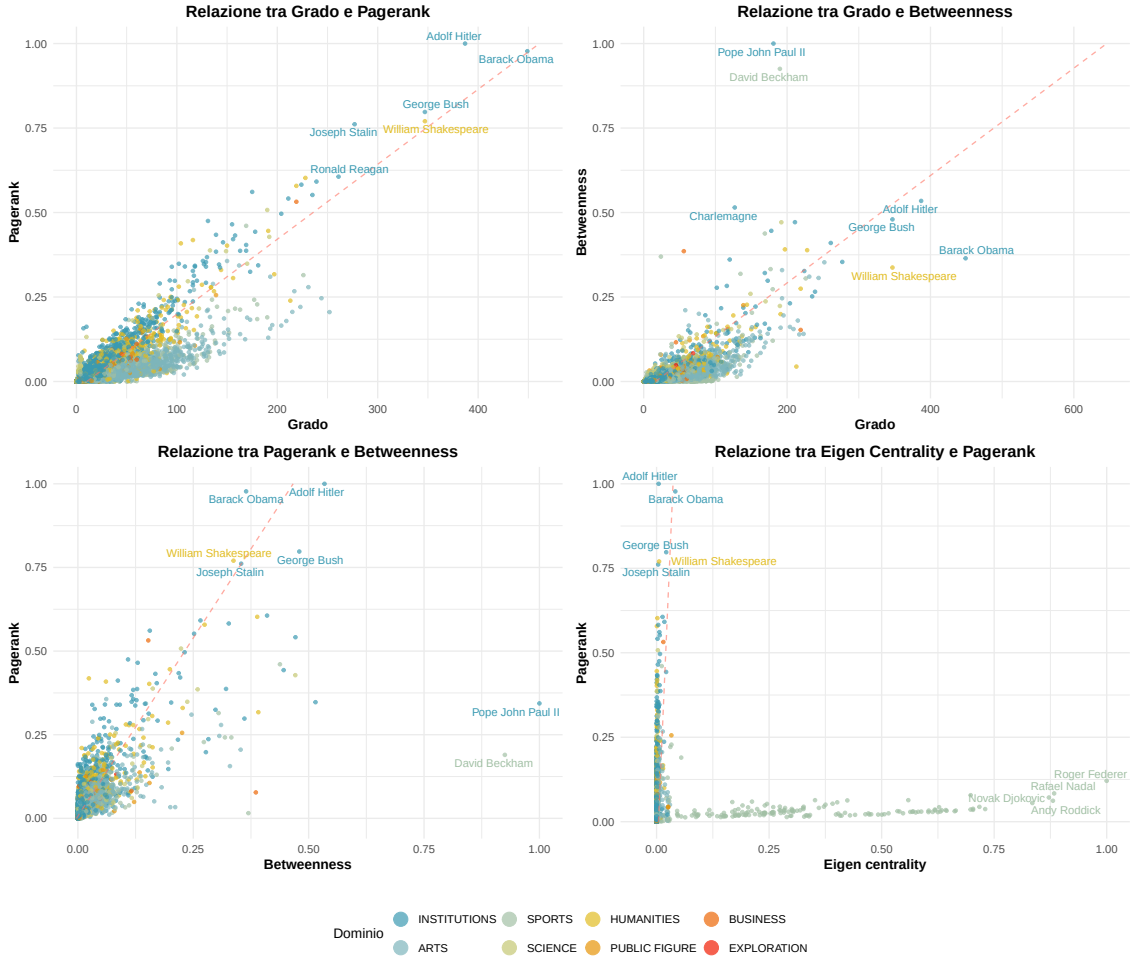


Figura 2.18: Relazione tra alcuni indici di centralità. In rosso, la prima componente principale per ogni coppia di indici

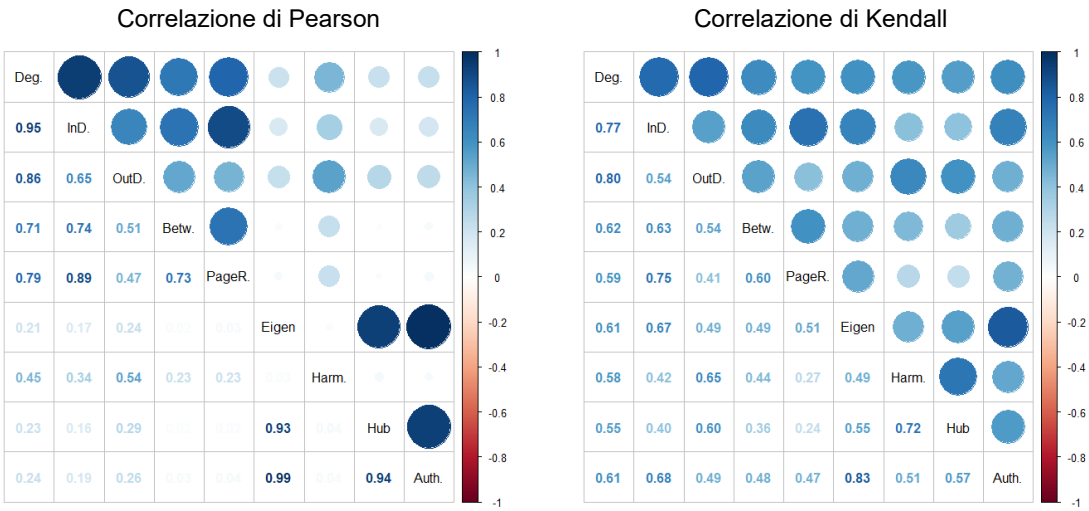


Figura 2.19: Correlazioni di Pearson e Kendall tra i vari indici di rilevanza

2.5 Visualizzazione della rete Pantheon

2.5.1 La rete intera

La visualizzazione grafica della rete può essere sfruttata per rilevare *pattern* e attori centrali in maniera qualitativa. A questo scopo, è stato utilizzato il software `Gephi` (Bastian *et al.*, 2009). In alcune rappresentazioni grafiche sono stati rimossi i nodi isolati, che non garantivano un'occupazione ottimale dello spazio da parte degli algoritmi di *layout*.

Un primo grafico, che sfrutta l'algoritmo HifanYu, permette di validare le ipotesi formulate in precedenza sulle interazioni tra individui provenienti da continenti diversi (Figura 2.20). I nordamericani sembrano infatti costituire un *cluster* a sé, e i pochi archi verso l'esterno coinvolgono personalità europee. Queste ultime, invece, sembrano più inclini a interagire con gli altri continenti, anche se risalta una certa autoreferenzialità. Gli altri continenti hanno apparizioni più sporadiche e meno concentrate, coerentemente con le analisi svolte in precedenza.

Un'indagine analoga sulle nazionalità dei nodi risulta molto più caotica, a causa del gran numero di modalità e quindi di colori nel grafico. Si evidenzia nuovamente l'omofilia, anche se in misura minore, coerentemente con le conclusioni della Sezione 2.4.2. Lo stesso discorso vale anche per i secoli di nascita: l'omofilia è spiccata, e gli attori sembrano disporsi lungo una retta temporale. Difatti, i nodi a sinistra nella rappresentazione grafica provengono da epoche più remote, mentre la metà destra del grafico è occupata in gran parte da individui del XX secolo. I grafici non sono riportati a causa della già citata natura confusionaria che li caratterizza.

L'osservazione della disposizione dei domini conferma che l'ambito d'interesse di un individuo è la caratteristica più determinante nella topologia della rete (Figura 2.21). La divisione in gruppi secondo il dominio d'appartenenza è infatti abbastanza netta: si confermano la grande autoaffinità delle figure artistiche e la loro connessione con le istituzioni e le discipline umanistiche. Appaiono meno centrali gli scienziati, anche se appartengono comunque al grande *cluster* centrale della rete. Gli sportivi si confermano estremamente autoreferenziali e, per certi versi, defilati rispetto agli individui degli altri domini. Si notano diversi *cluster* di figure sportive, che un'analisi più approfondita rivela coincidere con i vari sport praticati dagli individui del gruppo in questione.

Coerentemente con la numerosità nel *dataset*, la *community* più grande è costituita dai calciatori. Altri gruppi notevoli sono dati dai piloti e dai tennisti. La densità di quest'ultimo gruppo potrebbe spiegare la distorsione presente nell'indice di *eigen-centrality*, discussa nella Sezione 2.4.3. Inoltre, i due insiemi di nodi più vicini agli artisti sono dati da giocatori di basket e wrestler. Queste due categorie sono infatti più legate al mondo dello spettacolo, e in particolare a quello americano: ciò si riflette nella vicinanza, anche grafica, dei due *cluster* alle personalità del cinema.

Uno sguardo alla disposizione dei nodi in base al genere (Figura 2.22) è, nuovamente, coerente con le analisi quantitative svolte in precedenza: alcune sezioni della rete sono completamente prive di figure femminili, mentre altre presentano quasi parità di genere. Le aree caratterizzate da individui di epoche più remote o da ambiti professionali dominati dagli uomini, come le scienze o alcuni sport, appaiono

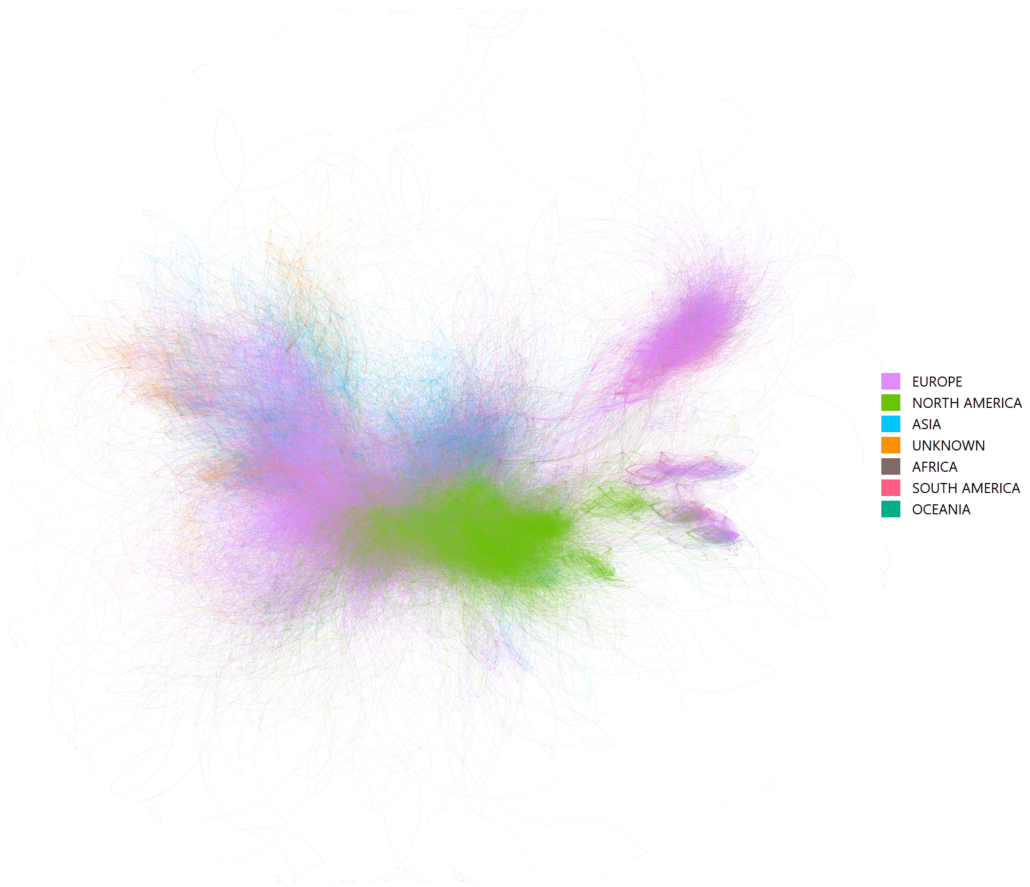


Figura 2.20: Grafico della rete Pantheon con l'algoritmo HifanYu. Il colore dei nodi indica il continente. Gli archi hanno il colore del nodo da cui escono.

monopolizzate dalle figure maschili. Un maggiore equilibrio si riscontra nel settore artistico, che presenta una significativa presenza femminile.

È interessante notare la presenza di un *cluster* distinto composto esclusivamente da donne: si tratta delle tenniste, i cui nodi sono attigui a quelli dei loro colleghi del circuito maschile. Questa configurazione suggerisce che i tennisti, indipendentemente dal genere, siano caratterizzati da un grado di omofilia superiore alla media, sia per quanto riguarda l'occupazione che il genere dei nodi.

È cruciale comprendere che gli algoritmi di *layout* operano esclusivamente sulla base della struttura topologica del grafo, ovvero sugli archi, senza considerare gli attributi intrinseci dei nodi stessi. La disposizione spaziale dei nodi è quindi determinata unicamente dalle relazioni tra gli elementi della rete, non dalle loro proprietà individuali. Solo successivamente si procede a esaminare la distribuzione degli attributi all'interno della visualizzazione ottenuta. La corrispondenza significativa tra i *cluster* e la distribuzione degli attributi dei nodi fornisce conferma empirica dell'ipotesi di omofilia.

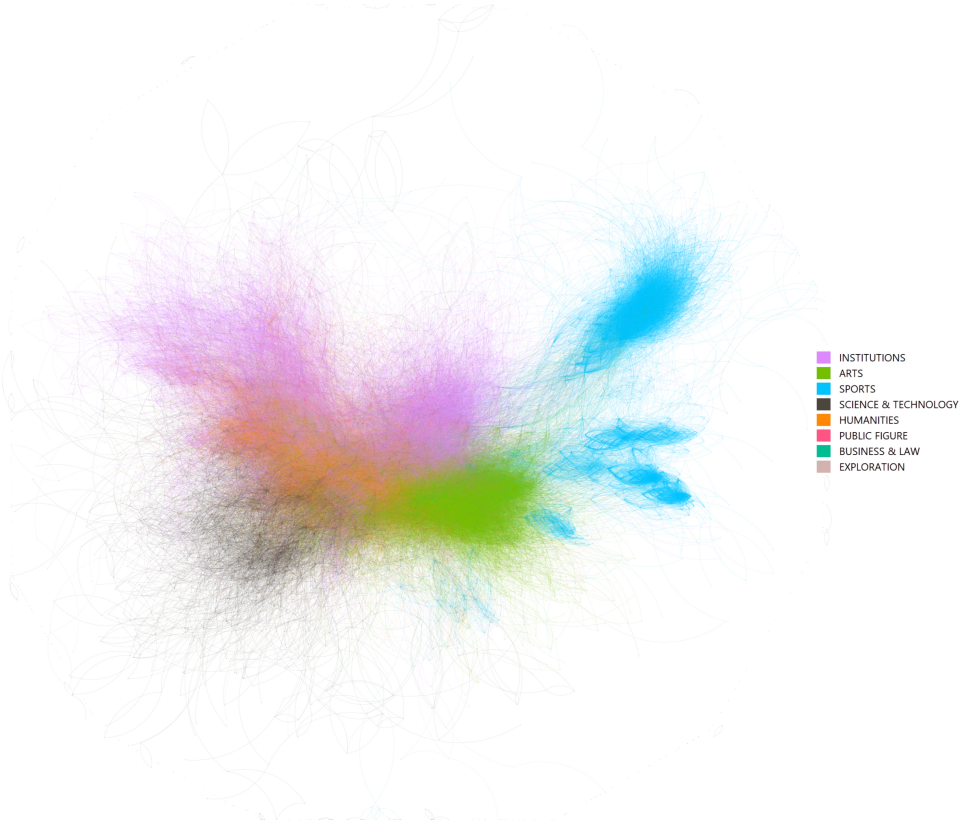


Figura 2.21: Grafico della rete Pantheon con l'algoritmo HifanYu. Il colore dei nodi indica il dominio. Gli archi hanno il colore del nodo da cui escono.

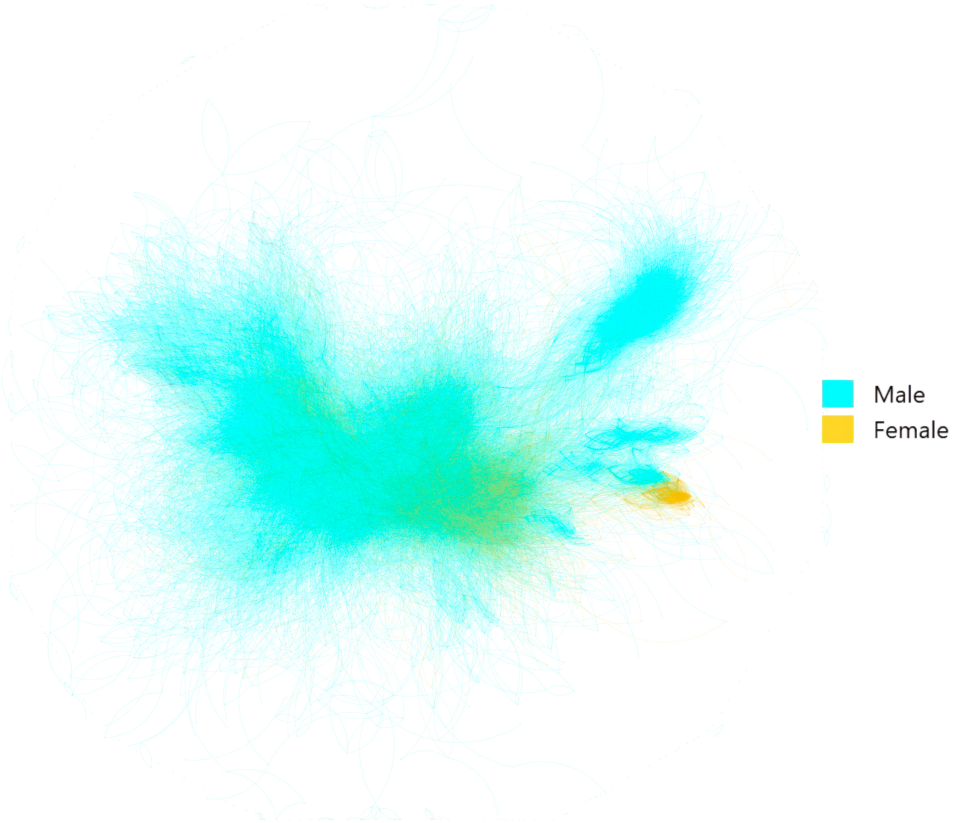


Figura 2.22: Grafico della rete Pantheon con l'algoritmo HifanYu. Il colore dei nodi indica il genere. Gli archi hanno il colore del nodo da cui escono.

2.5.2 Le sottoreti

È opportuno condurre anche un esame grafico approfondito sulle interazioni tra i diversi gruppi interni della rete Pantheon. Questa indagine si può svolgere considerando le singole modalità di alcune variabili categoriali come nodi di una rete con archi pesati. Difatti, le matrici Z mostrate nella Sezione 2.4.2 si rivelano matrici di adiacenza, le cui entrate corrispondono ai pesi degli archi. Le reti risultanti usano quindi come nodi i continenti, i Paesi, i secoli di nascita o i domini d'appartenenza, in maniera conforme a quanto presentato fin qui.

Il grafico dei continenti (Figura 2.23) conferma quanto già ipotizzato: l'Europa ha il ruolo più centrale, e questo emerge sia dalla sua posizione nel grafico, sia dal suo *score pagerank*, che la rende l'attore dalle dimensioni maggiori. Il Nord America è indubbiamente il secondo continente più rilevante; questi due continenti sono inoltre quelli di gran lunga più autoreferenziali. L'Asia emerge come terzo continente più centrale, ma in generale si può dire che tutti gli altri nodi sono periferici e dalla rilevanza ridotta, sia in termini di posizione che di dimensione. Osservazioni simili valgono per la rete dei Paesi (Figura 2.24): Stati Uniti e Regno Unito sono gli stati più centrali e autoreferenziali. Anche Italia, Francia e Germania risultano rilevanti, mentre le altre nazioni sembrano ricoprire un ruolo più marginale. Va notato come Paesi provenienti dallo stesso continente risultino tendenzialmente raggruppati. Inoltre, più una nazione è centrale del grafico, più, mediamente, sono rilevanti le dimensioni del nodo associato a essa: si evidenzia quindi un certo accordo tra l'algoritmo di *layout* e l'indice *pagerank*.

Più particolare è il risultato che l'algoritmo di *layout* propone per la rete dei secoli di nascita (Figura 2.25). Qui i nodi sono disposti in maniera abbastanza ordinata, con i secoli attigui che risultano più vicini. Inoltre, i secoli più recenti sono quelli più rappresentati e anche quelli più centrali e importanti. Risulta l'estrema autoreferenzialità del XX secolo, che però sembra essere penalizzato nello *score pagerank*: il 1800 ottiene infatti un punteggio più elevato. È plausibile che il XIX secolo funga da ponte tra il 1900 e gli altri secoli, e che questo ruolo sia tanto importante nella topologia della rete da renderlo più rilevante, secondo l'indice utilizzato.

Nella rete dei domini, vi è un grande *cluster* centrale dove si trova la maggior parte dei nodi (Figura 2.26). Risaltano, come prevedibile, le istituzioni, le arti e l'ambito umanistico. Quest'ultimo risulta inoltre avere uno *score pagerank* superiore a quanto la sua popolosità potesse suggerire. Il contrario è vero per gli sport: il dominio è molto popoloso, ma così autoreferenziale da risultare marginale nella rete. Le figure pubbliche e gli scienziati, nonostante una rilevanza ridotta, si trovano topologicamente vicini ai tre attori principali, evidenziando dei legami forti; più lontani sono invece le figure del *business* e soprattutto gli esploratori.

Si noti come i grafici delle sottoreti riportano risultati coerenti con quelli ottenuti dal grafico della rete intera, e con le analisi svolte sui flussi degli archi. In generale, la rete Pantheon è caratterizzata da grande autoaffinità, soprattutto all'interno dei domini. Il comportamento più peculiare è quello degli sportivi, che risultano numerosi quanto autoreferenziali e isolati. Del resto, non ci sono gruppi di nodi dai comportamenti anomali; anzi, i risultati sembrano in linea con quanto ci si potesse aspettare. Come già evidenziato, Europa e Nord America, con le nazioni che vi si

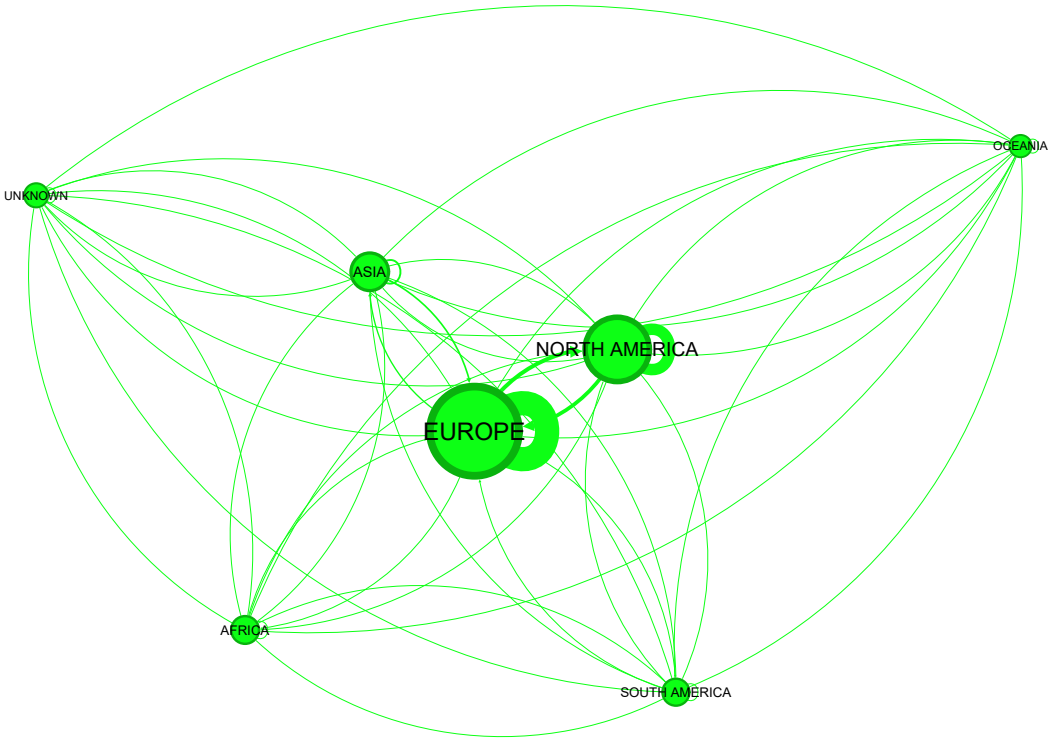


Figura 2.23: Grafico della rete pesata dei continenti con l'algoritmo ForceAtlas2. La dimensione dei nodi è proporzionale allo *score pagerank*, quella degli archi è proporzionale al peso.

trovano, emergono come le aree geografiche più rilevanti, mentre gli ultimi due secoli si confermano come il periodo storico dominante. Le figure che mediamente hanno più impatto sul mondo sono quelle istituzionali o artistico-umanistiche, aspetto che emerge chiaramente dalla rete Pantheon.

I risultati ottenuti sono quindi plausibili non solo alla luce dei dati a disposizione, ma anche considerando una visione generale dei secoli coperti dal *dataset*. Tuttavia, è importante notare che questa analisi potrebbe riflettere principalmente una prospettiva occidentale e quindi non catturare appieno l'influenza e l'importanza di figure provenienti da altre regioni del mondo, o nate in periodi storici a noi più remoti. La creazione di una nuova rete Pantheon a partire dai dati Pantheon 2.0 potrebbe mettere in luce dei *pattern* non rilevati in questo elaborato, o al contrario confermare e rafforzare queste conclusioni.

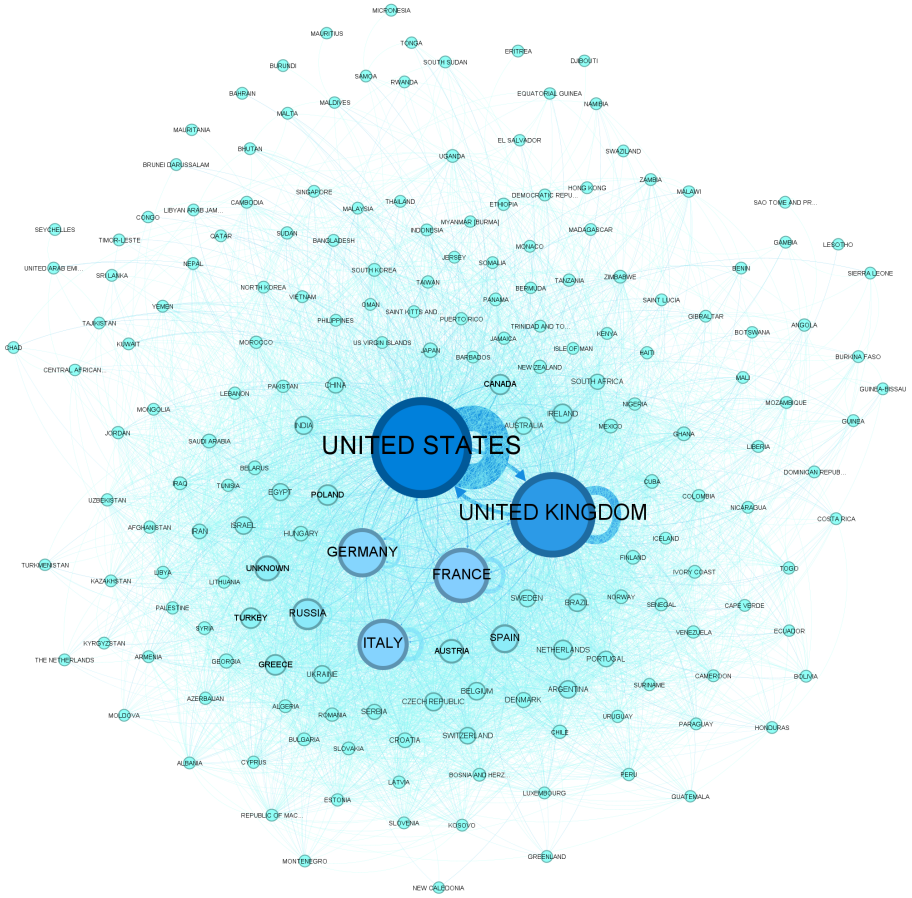


Figura 2.24: Grafico della rete pesata dei Paesi con l'algoritmo ForceAtlas2. La dimensione dei nodi è proporzionale allo *score pagerank*, quella degli archi è proporzionale al peso.

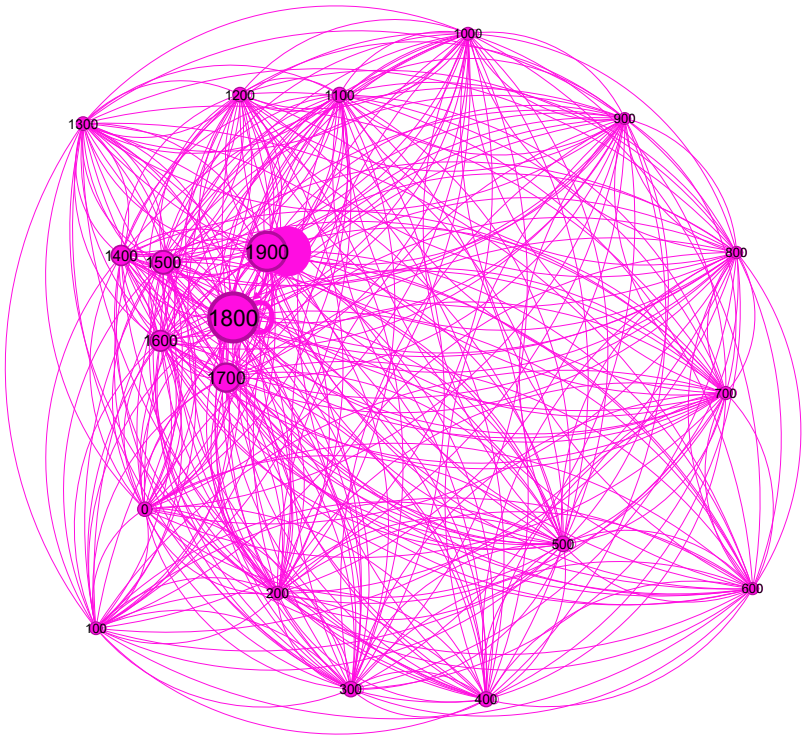


Figura 2.25: Grafico della rete pesata dei secoli con l'algoritmo ForceAtlas2. La dimensione dei nodi è proporzionale allo *score pagerank*, quella degli archi è proporzionale al peso.

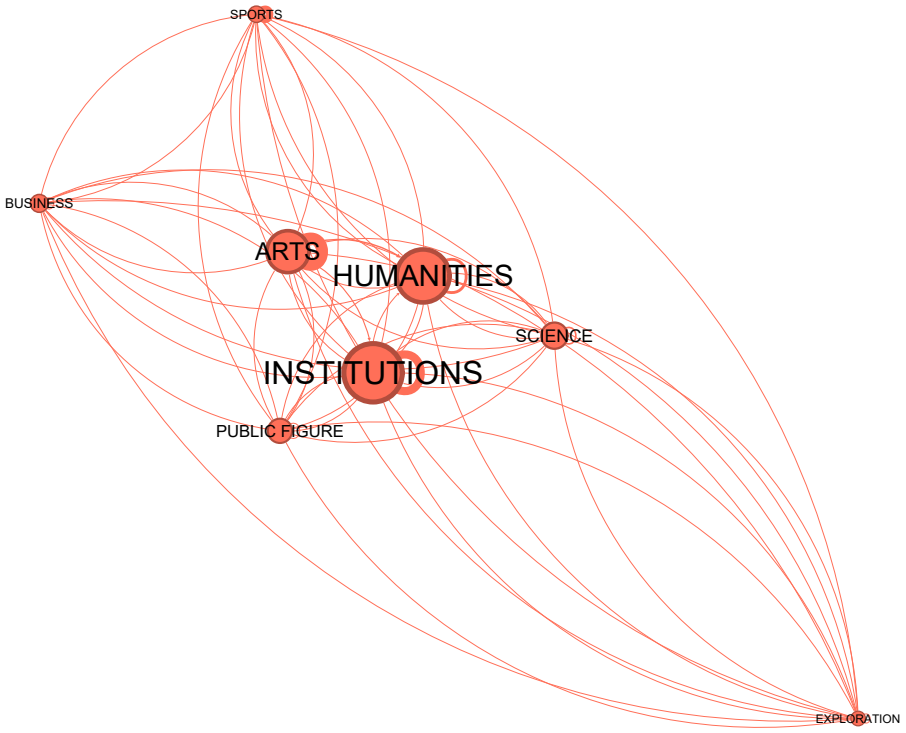


Figura 2.26: Grafico della rete pesata dei domini con l'algoritmo ForceAtlas2. La dimensione dei nodi è proporzionale allo *score pagerank*, quella degli archi è proporzionale al peso.

Capitolo 3

Analisi inferenziali delle strutture di rete

3.1 Alcuni modelli di rete

3.1.1 I modelli di Erdős-Rényi, *p1* e *p2*

La letteratura statistica offre una varietà di modelli, caratterizzati da diversi gradi di complessità e sofisticazione, per descrivere il comportamento di una rete. I modelli possono operare sia a livello locale, per l'identificazione di *cluster* o *communities* di attori, sia a livello globale, per individuare *pattern* rispetto ai quali gli attori interagiscono tra loro. In generale, la rete osservata è solo una delle possibili realizzazioni di interazioni tra i nodi, e l'obiettivo del modello non è spiegare i dati osservati nello specifico, ma descrivere il processo sottostante la generazione della rete in maniera più astratta e generalizzabile. Si assume infatti che vi sia un processo stocastico generatore della rete, e che gli archi che si vengono a formare abbiano o meno un certo livello di dipendenza da altri archi e dalle caratteristiche dei nodi. Per questo motivo, si pensa alla rete come a un sistema auto-organizzato di legami relazionali.

Viene generalmente assunta una funzione di verosimiglianza con un determinato numero di parametri, i quali sono poi stimati a partire dai dati. Di fatto, ci si trova a modellare le $N(N-1)$ entrate della matrice di adiacenza $\mathbf{Y}$ (qui si esclude la diagonale), che vengono trattate come variabili casuali bernoulliane. Ciò rende evidente come il numero di parametri possa variare in un intervallo molto esteso; come accade spesso in statistica, l'obiettivo è individuare e stimare un modello in grado di cogliere i *pattern* in maniera soddisfacente, ma tralasciando il rumore. Va da sé che un modello con un solo parametro, o con $N(N-1)$ parametri, non descriveranno adeguatamente il meccanismo di generazione dei dati, rispettivamente a causa di *underfitting* e *overfitting*.

Il primo, e più semplice, modello proposto è quello di Erdős-Rényi (Erdős e Rényi, 1959). Indicata con $\mathbf{Y} = \{Y_{ij}\}$ la matrice di adiacenza associata alla rete, l'assunzione fondamentale è di indipendenza e identica distribuzione (i.i.d.) fra le variabili Y_{ij} che denotano la presenza o meno di un link tra due attori. Posto $P(Y_{ij} = 1) = \theta$ e, in modo complementare, $P(Y_{ij} = 0) = 1 - \theta$, la probabilità di osservare una particolare

rete è quindi data da:

$$\Pr(\mathbf{Y} \mid \theta) = \prod_{i,j} \theta^{Y_{ij}} (1-\theta)^{1-Y_{ij}} \tag{3.1}$$

Presentando un solo parametro (θ), insieme all'assunzione di indipendenza ed equiprobabilità, è intuibile come il modello non sia in grado di cogliere la complessità nelle interazioni tra attori. Difatti, il modello di Erdős-Rényi viene utilizzato come modello nullo, ovvero uno in cui non vi è struttura.

Un modello più complesso, per le reti dirette, considera i 4 possibili stati di un arco tra i nodi n_i e n_j: assente, reciproco o presente solo in una o l'altra direzione. Il modello, detto *p1* (Holland e Leinhardt, 1981), si presenta, in forma semplificata e più intuitiva, come segue:

$$\Pr(\mathbf{Y} = y) \propto \exp\left(\theta \sum_{i,j} Y_{ij} + \sum_i \alpha_i \sum_j Y_{ij} + \sum_j \beta_j \sum_i Y_{ij} + \rho m\right)$$

dove $m = \sum_{i<j} Y_{ij} Y_{ji}$ è il numero di archi reciproci. È presente ancora un parametro (θ) che indica la tendenza generale nella rete alla formazione di archi, come nella formula (3.1). Si aggiungono però 3 ulteriori parametri: α, vettore $N \times 1$ con coefficienti di produttività specifici per ogni nodo; β, vettore $N \times 1$ con coefficienti di attrattività caratteristici di ogni nodo; e infine lo scalare ρ, a indicare la tendenza alla reciprocità. Imponendo l'identificabilità del modello, si ottengono $2n$ parametri: è evidente come questo valore possa crescere rapidamente in grandi reti. Inoltre, si può trasformare ρ in un vettore $N \times 1$, in modo da rendere la reciprocità una caratteristica specifica di ogni nodo, e non più omogenea su tutta la rete; è però necessario imporre ulteriori restrizioni per evitare la non-identificabilità del modello.

Un'estensione di questo modello è il *p2* (Van Duijn *et al.*, 2004). In questo caso, i coefficienti α e β sono tratti da delle distribuzioni a media nulla, e i coefficienti di densità e reciprocità diventano dipendenti dalla diade (coppia di nodi) a cui si riferiscono. Ciò comporta una maggiore parsimonia nel numero di parametri, e rende *p2* un modello a effetti casuali. Inoltre, la peculiarità di questo modello è la possibile presenza di covariate nelle distribuzioni degli effetti casuali, che si aggiungono per mezzo di 2 parametri per ogni attributo di nodo rilevato. Gli autori evidenziano come sia necessario prestare attenzione alla collinearità tra i parametri da stimare, poiché questo fenomeno non è sempre ovvio da individuare e può portare ai problemi di non identificabilità già noti in statistica.

3.1.2 Modelli a blocchi e a variabili latenti

Un approccio differente alla modellazione delle reti è dato dai modelli a blocchi (*block model*). In questo caso, i nodi sono mappati in dei blocchi (o gruppi) e si assume che la distribuzione dell'interazione Y_{ij} tra i nodi n_i e n_j dipenda solo dai blocchi di appartenenza dei due nodi. La separazione dei nodi in gruppi può avvenire *a priori*, se si dispone di particolari informazioni o attributi rilevati. Il risultato è facilmente interpretabile, anche se abbastanza rigido; oltretutto, la scelta del numero di blocchi non è sempre univoca e immediata.

Un'ulteriore classe di modelli presente in letteratura è costituita dai modelli a variabili latenti. In questo caso si assume che i dati siano dipendenti da un insieme di

variabili non osservabili, o latenti. Sono proprio queste variabili a essere oggetto della modellazione; la struttura appare quindi gerarchica, e l'inferenza si serve spesso di strumenti bayesiani. Hoff *et al.* (2002) hanno proposto modelli basati sulla distanza o sull'angolo formato da due nodi negli spazi latenti; altri autori hanno presentato modelli di *clustering* a posizioni latenti.

Rientra nei modelli a variabili latenti anche il modello stocastico a blocchi (*stochastic block model*): i *cluster* di appartenenza dei nodi sono infatti trattati come variabili latenti su cui fare inferenza. Un'estensione di questo modello permette inoltre di rendere variabile la posizione del nodo n_i in base al nodo n_j con cui sta interagendo: si parla di appartenenza mista (*mixed membership stochastic block model*).

Una rassegna dei principali modelli, tra cui quelli appena presentati, con relativi riferimenti, si può trovare in Salter-Townshend *et al.* (2012).

3.2 Gli ERGM

3.2.1 Generalità

Un grande limite dei modelli Erdős-Rényi, *p1* e *p2* è l'assunzione di indipendenza tra gli archi e quindi tra le diadi (*dyad-independence*). Questa configurazione rende infatti agevoli la stima e l'interpretazione, ma allo stesso tempo troppo rigida e spesso poco realistica la modellazione risultante. Una classe che permette di assumere dipendenze più flessibili nella topologia della rete è quella dei modelli della famiglia esponenziale per grafi casuali (*Exponential family Random Graph Models*, o ERGM), anche detti modelli p^*. La probabilità di osservare una certa rete è data dalla formula:

$$\Pr(\mathbf{Y};\theta) = \frac{\exp\{\theta^\top g(\mathbf{Y})\}}{\kappa(\theta)} \tag{3.2}$$

dove $\theta \in \Theta \subset \mathbb{R}^p$ è un vettore di p parametri, $g(\mathbf{Y}) \in \mathbb{R}^p$ è un vettore di p statistiche di rete e $\kappa(\theta) = \sum_{Z \in \mathcal{Y}} \exp\{\theta^\top g(\mathbf{Z})\}$ è una costante di normalizzazione calcolata sull'insieme $\mathcal{Y}$ di tutte le possibili reti. È evidente come questo insieme abbia una cardinalità estremamente elevata, che rende impossibile un approccio enumerativo in fase di stima. Per una rete diretta, $\mathcal{Y}$ conta infatti $2^{N(N-1)}$ elementi. Le statistiche di rete $g(\mathbf{Y})$ vengono scelte dall'analista e possono includere sia quantità relative agli attributi di nodo, sia valori che rappresentano una tendenza generale nella rete, come la densità o la transitività.

I modelli Erdős-Rényi, *p1* e *p2* sono esplicitabili secondo la formula (3.2) e sono quindi da considerarsi casi speciali degli ERGM. In particolare, rientrano nel gruppo di ERGM che in letteratura sono stati chiamati modelli logistici, poiché la loro stima e interpretazione sono analoghe a quelle di una regressione logistica. Ciò non è vero in generale per gli ERGM se si fa cadere l'assunzione di indipendenza tra diadi, come è usuale.

Si distinguono tre possibili ipotesi di dipendenza locale per un ERGM: indipendenza diadica, dipendenza markoviana e dipendenza parziale condizionata. Nel caso di indipendenza diadica, come già osservato, il valore assunto dall'arco Y_{ij} è stocasticamente indipendente da tutti gli altri, e la funzione di densità della rete altro

non è che il prodotto delle singole densità. La dipendenza markoviana, introdotta da Frank e Strauss (1986), implica dipendenza tra due archi se essi sono incidenti su un attore comune; l'arco Y_{ij} dipende quindi solo dagli archi incidenti su n_i e n_j. Gli archi che non condividono nodi sono indipendenti, ma solo condizionatamente al resto della rete. È infatti evidente come, secondo la definizione appena data, tra gli archi Y_{ij} e Y_{kl} non ci sia dipendenza markoviana; tuttavia, l'esistenza dell'arco Y_{jk} fa sì che esso agisca da mediatore e generi una dipendenza indiretta tra Y_{ij} e Y_{kl}. Infine, la dipendenza parziale condizionata (Pattison e Robins, 2002) estende il concetto di località oltre gli archi incidenti su nodi comuni, cosicché possa esserci dipendenza tra Y_{ij} e Y_{kl}, se una terza parte (ad esempio, Y_{jk}) li collega. Tale dipendenza non si esaurisce nella mediazione di Y_{jk}, come nel caso markoviano, ma si aggiunge a essa, e permette di modellare in modo più esteso ed esaustivo. L'ipotesi sulla forma della dipendenza da adottare può basarsi su conoscenze teoriche pregresse sul fenomeno in esame, ma deve essere successivamente validata attraverso l'adattamento del modello ai dati osservati.

Le diverse dipendenze che sono state presentate vengono incluse nella formula del modello tramite le varie statistiche di rete che costituiscono il vettore $g(\mathbf{Y})$. La definizione e il significato delle statistiche possono cambiare per reti dirette o indirette; talvolta, una statistica per reti dirette non è utilizzabile per reti indirette, e viceversa. Alcuni esempi di statistiche di rete sono dati da:

- $L(\mathbf{Y}) = \sum_{i,j} y_{ij}$
- $T(\mathbf{Y}) = \sum_{i<j<m} y_{ij} y_{jm} y_{mi}$ (per reti indirette)
- $S_k(\mathbf{Y}) = \sum_i \binom{D(n_i)}{k}$
- $D_k(\mathbf{Y}) = \sum_i \mathbb{1}\left(D(n_i) = k\right)$
- $U_v(\mathbf{Y}) = \sum_{i,j} y_{ij} \cdot \mathbb{1}(x_{vi} = x_{vj})$

dove $\mathbb{1}(\cdot)$ è la funzione indicatrice, pari a 1 se l'argomento è vero e 0 altrimenti.

$L(\mathbf{Y})$ rappresenta il numero totale di archi nella rete ed è equivalente alla densità (equazione 2.1) moltiplicata per $N(N-1)$. $T(\mathbf{Y})$ è il numero di triangoli ed è una delle statistiche utilizzate per la transitività; la sua definizione per le reti dirette è più complessa, ma mantiene un'interpretazione simile. $S_k(\mathbf{Y})$ è il numero di k-stelle (*k-star*), ovvero configurazioni in cui un nodo centrale è connesso a k altri nodi. $D_k(\mathbf{Y})$ è la distribuzione del grado nella rete, mentre $U_v(\mathbf{Y})$ è la statistica di omofilia per l'attributo di nodo v. Per una rassegna di statistiche più esaustiva si rimanda a Goodreau (2007).

Si nota quindi che vi possono essere statistiche che descrivono tendenze dei singoli nodi o generali nella rete, e che queste sono anche legate agli attributi di nodo. È facile far aumentare copiosamente il numero di parametri di un modello, e allo stesso tempo risulta complesso capire quali statistiche includere. Difatti, una rete molto densa presenterà naturalmente elevata transitività e quindi un alto numero di triangoli; è quindi possibile che il risultato in fase di stima sia un parametro grande per $L(\mathbf{Y})$ e pressoché nullo per $T(\mathbf{Y})$. Un altro fenomeno che si verifica in maniera talvolta imprevedibile è la dipendenza lineare: si noti come la statistica

$S_1(\mathbf{Y})$ sia uguale a $L(\mathbf{Y})$. È quindi opportuno prestare attenzione alle statistiche che si inseriscono nel modello per evitare problemi di non-identificabilità.

Le interpretazioni delle statistiche e dei relativi parametri non sono sempre chiare, e lo stesso si può dire delle interazioni tra diverse statistiche di rete. Inoltre, inserire variabili di interazione aumenta considerevolmente la complessità del modello, a scapito della comprensibilità dello stesso; per questo motivo, sono incluse raramente. È quindi usuale supporre quali parametri possano risultare rilevanti, per poi procedere stimando iterativamente finché il processo di *trial and error* non sembra portare a una conclusione valida.

A dispetto della grande flessibilità, spesso la stima degli ERGM è problematica (Handcock, 2003a,b), ed empiricamente si è riscontrato che l'inclusione di determinate statistiche nel modello amplifica questo problema; un esempio è il numero di triangoli. Sono state proposte diverse statistiche più elaborate per ovviare a questi problemi, e tra quelle che hanno riscosso più successo vi sono quantità che pesano geometricamente la quantità di triangoli o stelle della rete, la distribuzione del grado o il numero di nodi adiacenti a una data diade. Una forma generalmente usata è la seguente:

$$gwg(\mathbf{Y} \mid \tau) = e^{\tau} \sum_{k} \{1 - (1 - e^{-\tau})^k\} g(\mathbf{Y})$$

dove si antepone *gw* (*geometrically weighted*) al nome della statistica, la generica $g(\mathbf{Y})$ in questo caso. k è invece l'ordine della statistica, e può essere, ad esempio, il valore k delle *k-star* o il grado dei nodi; τ è invece il parametro di decadimento (*decay parameter*), fissato arbitrariamente.

Il risultato è una statistica singola che sintetizza un'ampia gamma di interazioni complesse. Questa sintesi permette di rappresentare fenomeni che altrimenti richiederebbero numerosi parametri, peraltro spesso discordanti tra loro a causa della forte collinearità. Tale riduzione offre una visione d'insieme più accessibile e interpretabile, pur sacrificando talvolta la precisione. Per le reti dirette, queste statistiche, ove possibile, vanno considerate sia in direzione *in* che *out*: la maggior complessità della rete si riflette infatti nel maggior numero di parametri necessari. La modellazione delle reti dirette è quindi più difficoltosa, sia per la difficoltà nell'individuare statistiche adeguate, sia per il carico computazionale che aumenta sensibilmente.

L'uso di statistiche pesate geometricamente senza un parametro di decadimento τ prefissato, ma variabile, fa sì che la distribuzione del modello appartenga alla famiglia esponenziale curva, e presenti una serie di altri vantaggi e svantaggi che esulano dagli obiettivi di questo elaborato. Per un approfondimento, si rimanda a Hunter (2007).

3.2.2 Stima e bontà di adattamento

In linea teorica, la specifica del modello statistico parametrico (3.2) consente di stimare i parametri con i metodi della massima verosimiglianza. Tuttavia, il principale problema è rappresentato dal termine $\kappa(\theta)$ dell'equazione (3.2): come già accennato, enumerare tutte le possibili reti è impraticabile se il numero di nodi è superiore a 30. Esistono diverse proposte che tentano di risolvere o aggirare questo problema, ma la ricerca deve ancora fare dei passi avanti.

L'intrattabilità della funzione di verosimiglianza ha portato ad approcci alternativi alla semplice ottimizzazione, come la stima di massima verosimiglianza Monte Carlo

(*Monte Carlo Maximum Likelihood Estimate*, o MCMLE) o la stima di massima verosimiglianza con Catena di Markov Monte Carlo (*Markov Chain Monte Carlo Maximum Likelihood Estimate*, o MCMCMLE), basate sull'idea di approssimare la verosimiglianza mediante simulazione. Un metodo differente, spesso affiancato a quelli appena menzionati, è l'uso della pseudo-verosimiglianza (*Maximum Pseudo-Likelihood Estimate*, o MPLE), che sfrutta l'uso delle cosiddette statistiche di variazione, o *change statistic*. Queste rappresentano la variazione nelle statistiche della rete quando un certo arco y_{ij} cambia il suo valore da 0 a 1 o viceversa, fermo restando il resto della rete. Si sfrutta poi la forma della densità degli ERGM per ottenere delle log-quote (*log-odds*) e la loro relazione con i parametri, in maniera analoga alla regressione logistica:

$$\log\left(\frac{\Pr\left(\mathbf{Y}_{ij}=1 \mid y_{ij}^{c}\right)}{\Pr\left(\mathbf{Y}_{ij}=0 \mid y_{ij}^{c}\right)}\right)=\theta^{\top} d_{ij}(y)$$

dove y_{ij}^{c} corrisponde a tutte le entrate di $\mathbf{Y}$ tranne y_{ij} e $d_{ij}(y)$ è il vettore delle statistiche di variazione per l'arco y_{ij}. La log-quota per uno specifico arco è quindi data da θ moltiplicato per il cambiamento osservato nelle statistiche di rete cambiando il valore dell'arco in questione da 1 a 0, o viceversa. La comprensione delle proprietà delle pseudo-verosimiglianze è alquanto limitata, e spesso esse non sembrano allinearsi con quelle desiderate, ovvero quelle della vera funzione di verosimiglianza. Inoltre, la regressione logistica assume osservazioni indipendenti; l'assunzione è evidentemente errata per dati di rete, e ciò comporta che più vi è dipendenza nella rete, più aumenta la distorsione delle stime di pseudo-verosimiglianza. Per questo motivo, sono più diffusi gli approcci Monte Carlo già menzionati, che però presentano limiti notevoli nell'ordine delle reti che sono in grado di trattare ($N < 1000$).

La grande flessibilità di questa classe di modelli dovrebbe permettere di fare inferenza in molteplici situazioni. In realtà, però, questo non è sempre vero: una difficoltà che si presenta frequentemente è la degenerazione del modello (*degeneracy,* Handcock, 2003a,b). In breve, questo fenomeno si verifica quando una specificazione del modello apparentemente ragionevole implica in realtà che una grande massa di probabilità della funzione di densità si concentri su poche possibili reti. La rete risultante da modelli degeneri, o quasi, è spesso "piena" o "vuota" ($y_{ij} = 0$ o $y_{ij} = 1$, $\forall\ i, j$). Quando la catena di Markov si trova in questa situazione, le stime dei parametri tendono a infinito e, di fatto, la MCMCMLE non esiste. L'uso della pseudo-verosimiglianza offre invece risultati più stabili; non si tratta di una soluzione al problema, ma solo di un modo per ignorarlo: se la specificazione del modello è errata, non è certo la convergenza di un algoritmo di stima che sfrutta approssimazioni a rendere valide le conclusioni.

L'inclusione di statistiche markoviane nel modello tende a favorire la *degeneracy*. L'aggiunta o la rimozione di un arco nella rete può avere un'influenza importante sulle *change statistics* delle quantità markoviane, soprattutto se l'arco in questione ha una funzione di ponte all'interno della rete; altre statistiche sono invece molto meno sensibili a variazioni minime della topologia del sistema. Le variazioni nelle statistiche markoviane generano e amplificano le dipendenze tra gli archi in un *feedback loop* che porta il modello a degenerare. Si ritiene che lo spazio parametrico effettivo Θ sia in realtà un sottoinsieme molto ridotto di quello teorico ($\mathbb{R}^p$), e che la *degeneracy*

sia dovuta alla ricerca delle stime di θ fuori dallo spazio parametrico effettivo. Le statistiche pesate geometricamente, menzionate in precedenza, aiutano a restringere lo spazio di ricerca, e a ridurre il rischio di degenerazione.

La stima dei modelli p^* risulta dunque difficoltosa sia per ragioni computazionali che statistiche. Nuovi approcci che sembrano più promettenti sono basati su *bootstrap*, prospettive bayesiane o algoritmi MCMC a troncamento variabile. Un'esposizione estesa ed esaustiva dei metodi di stima per ERGM richiederebbe una trattazione a parte, ed esula dagli obiettivi di questo elaborato. Si rimanda a van Der Pol (2019) per maggiori approfondimenti.

Un'altra complicazione presentata da questi modelli di rete è la valutazione della bontà d'adattamento. Un primo requisito, necessario ma non sufficiente perché il modello sia valido, è la convergenza dell'algoritmo verso un modello non degenere. Inoltre, nel caso di indipendenza diadica, il calcolo della pseudo-verosimiglianza è auspicabilmente poco distorto e quindi l'uso di criteri di informazione, quali l'AIC, è possibile. Anche i metodi Monte Carlo forniscono un valore per la verosimiglianza, ma questo è approssimato e, come già accennato, tanto distorto quanto più è presente dipendenza tra archi nel modello.

Di conseguenza, il metodo più diffuso per la bontà d'adattamento si basa sulla simulazione (Hunter *et al.*, 2008b). In pratica, si simula un certo numero di reti dalla densità stimata, con una procedura MCMC analoga a quella di stima. Si calcolano poi una serie di statistiche di rete, anche di ordine elevato (relative alle interazioni tra più nodi), sulla rete originale e su quelle stimate, e si procede a un confronto grafico; la simulazione permette anche di costruire test d'ipotesi con relativi *p-value* e intervalli di confidenza. In genere, si raffigurano dei *boxplot* per la distribuzione delle statistiche ottenute dalle reti simulate, con i relativi intervalli di confidenza per le medie, e si esamina se e quanto tali statistiche siano distanti dai valori osservati nella rete originale. Questa procedura si realizza per più statistiche, e in particolare su quantità non presenti nel modello: è infatti plausibile che le statistiche incluse nel modello siano descritte in maniera soddisfacente dallo stesso (anche se ciò non è sempre vero); quello che ci si chiede è se il modello sia in grado di rappresentare sufficientemente bene anche dei comportamenti che non sono esplicitamente presenti nella sua formulazione. Nel confronto tra le topologie della rete osservata e di quelle simulate dal modello stimato, è quindi usuale servirsi di statistiche che considerano diverse prospettive: globale, locale e di transitività. Un approccio multidimensionale garantisce un confronto più accurato e completo su più livelli. Si rimanda a Hunter *et al.* (2008b) per approfondimenti.

3.3 ERGM su alcune sottoreti di Pantheon

3.3.1 Premesse

In questo paragrafo verranno presentati i risultati relativi alle stime e alle valutazioni delle bontà di adattamento di alcuni dei modelli descritti. Come anticipato, la dimensione e l'ordine della rete Pantheon precludono la possibilità di adattare un modello all'intera rete. Per questo motivo, sono state selezionate, a titolo di esempio, alcune sottoreti aventi caratteristiche e dimensioni diverse, nonché un modesto livello

di sparsità. Questo ha consentito di ridurre i problemi di *degeneracy*, oltre che mantenere uno sforzo computazionale gestibile, perché, come vedremo, anche in questo modo la stima non è stata esente da problemi.

Per i modelli che seguono, al fine di evitare dipendenze lineari tra i parametri, si sono poste come categorie di riferimento *Arts*, *Africa* e *Female*, rispettivamente per i domini (*domain*), i continenti (*continent*) e il genere (*gender*). Nelle simulazioni per la bontà di adattamento si sono usati campioni di 500 reti e un *burnin* di 1000 per la procedura MCMC. Per le analisi di questa sezione sono stati utilizzati in `R` i pacchetti della suite `statnet` (Handcock *et al.*, 2008; Goodreau *et al.*, 2008), tra cui le librerie `ergm` (Hunter *et al.*, 2008a; Handcock *et al.*, 2023; Krivitsky *et al.*, 2023) e `network` (Butts, 2008, 2015).

3.3.2 Un modello di indipendenza diadica

La prima sottorete presa in analisi è costituita dai 525 attori nati nel XVIII secolo. Il modello stimato è di indipendenza diadica, e tiene quindi conto solo delle caratteristiche dei nodi per descrivere la rete. Questa formulazione ha permesso di modellare una rete piuttosto grande, e allo stesso tempo garantisce una facile interpretazione dei parametri. Questi infatti si interpretano in maniera analoga alla regressione logistica, quindi come log-rapporti di quote. Formalmente, usando la statistica di omofilia $U_v(y)$ per la variabile categoriale v come esempio:

$$\log\left(\frac{\frac{\Pr(y_{ij}=1|U_v(y_{ij})=1)}{\Pr(y_{ij}=0|U_v(y_{ij})=1)}}{\frac{\Pr(y_{ij}=1|U_v(y_{ij})=0)}{\Pr(y_{ij}=0|U_v(y_{ij})=0)}}\right) = \theta_v$$

I risultati sono abbastanza in linea con le ipotesi che si erano avanzate nelle analisi esplorative (Capitolo 2): nodi che condividono lo stesso dominio, genere o luogo di nascita sono più inclini a creare un legame (Tabella 3.1). La tendenza generale a creare legami è piuttosto bassa, e sembra essere inferiore a quella degli artisti solo per le figure umanistiche (non in maniera significativa) e scientifiche; nodi appartenenti ad altri domini creano più archi, ma solo le figure pubbliche in modo significativo al 5%. I nati di ogni continente sembrano più attivi nel creare connessioni rispetto agli africani, ma nessuno in maniera significativa. Non risultano nemmeno differenze statisticamente significative tra uomini e donne; questa sottorete ha però una presenza minima di figure femminili (circa il 6%), quindi non sembrano esserci le basi per trarre conclusioni in questo senso. Una rimozione dei termini semplificherebbe il modello, ma gli ERGM non sono sempre robusti rispetto alla rimozione di alcune variabili; una riduzione del modello può infatti portare a problemi di *degeneracy* o di mancata convergenza. In questo caso, le variabili non significative sono state tenute anche per motivi interpretativi.

Come ci si può aspettare, nonostante i requisiti computazionali (convergenza e non-*degeneracy*) siano stati soddisfatti, l'adattamento del modello non è generalmente ottimale. L'assunzione di indipendenza diadica si rivela infatti quantomeno dubbia, e il modello non sembra catturare in maniera soddisfacente il meccanismo generatore della rete. Ciò si riflette nella distribuzione di alcune statistiche di rete, nodali e locali, rilevate sulle reti simulate dal modello stimato (Figura 3.1). Tali statistiche

	Stima	**Std. Error**	**Pr(>|z|)**
$edges$	-8.90	1.43	0.00
U_{domain}	1.82	0.04	0.00
$U_{continentname}$	1.65	0.08	0.00
U_{gender}	0.40	0.10	0.00
$U_{countryname}$	1.66	0.04	0.00
$views$	$6 \cdot 10^{-6}$	$1.6 \cdot 10^{-7}$	0.00
$domain_{Business}$	0.58	0.31	0.06
$domain_{Exploration}$	0.10	0.13	0.42
$domain_{Humanities}$	-0.04	0.04	0.37
$domain_{Institutions}$	0.05	0.04	0.19
$domain_{PublicFigure}$	0.76	0.08	0.00
$domain_{Science}$	-0.21	0.04	0.00
$continent_{Asia}$	0.73	0.72	0.31
$continent_{Europe}$	0.24	0.71	0.74
$continent_{NorthAmerica}$	0.50	0.71	0.48
$continent_{SouthAmerica}$	0.69	0.72	0.34
$continent_{Unknown}$	0.85	0.78	0.28
$gender_{Male}$	0.06	0.10	0.52

Tabella 3.1: Valore dei coefficienti, con relativi *standard error* e *p-value*, del modello di indipendenza diadica stimato per i nati nel XXVIII secolo

presentano una media osservata sensibilmente inferiore alla mediana, portando all'ipotesi che poche reti simulate siano degli *outlier* che rendono distorto lo stimatore della media, notoriamente poco robusto, aumentandone la varianza e quindi l'ampiezza degli intervalli di confidenza. Le statistiche osservate sono sempre molto vicine al limite inferiore di questi intervalli, ma data la natura della distribuzione e dei test multipli che si stanno effettuando sembra ragionevole concludere che l'adattamento è insoddisfacente. Un confronto tra la rete osservata e una rete simulata risulta poco pratico a causa delle dimensioni della rete, ma lo scarso adattamento mostrato dalle verifiche grafiche sembra sufficiente per sostenere l'ipotesi di inadeguatezza del modello.

3.3.3 La rete dei personaggi irlandesi

Un'altra sottorete che si può analizzare è costituita dai 53 nodi di nazionalità irlandese. Le dimensioni ridotte permettono di stimare un modello più complesso, che vada oltre l'indipendenza diadica. Oltre ad alcuni attributi di nodo e statistiche di omofilia, vengono considerati la reciprocità degli archi, il numero di connessioni mediate da un terzo nodo (*twopath*) e il numero di *partner* condivisi da due nodi collegati. Quest'ultima statistica è pesata geometricamente, con $\tau = 0.25$, da cui il nome del parametro *gwesp* (*geometrically weighted edge-wise shared partners*). È evidente

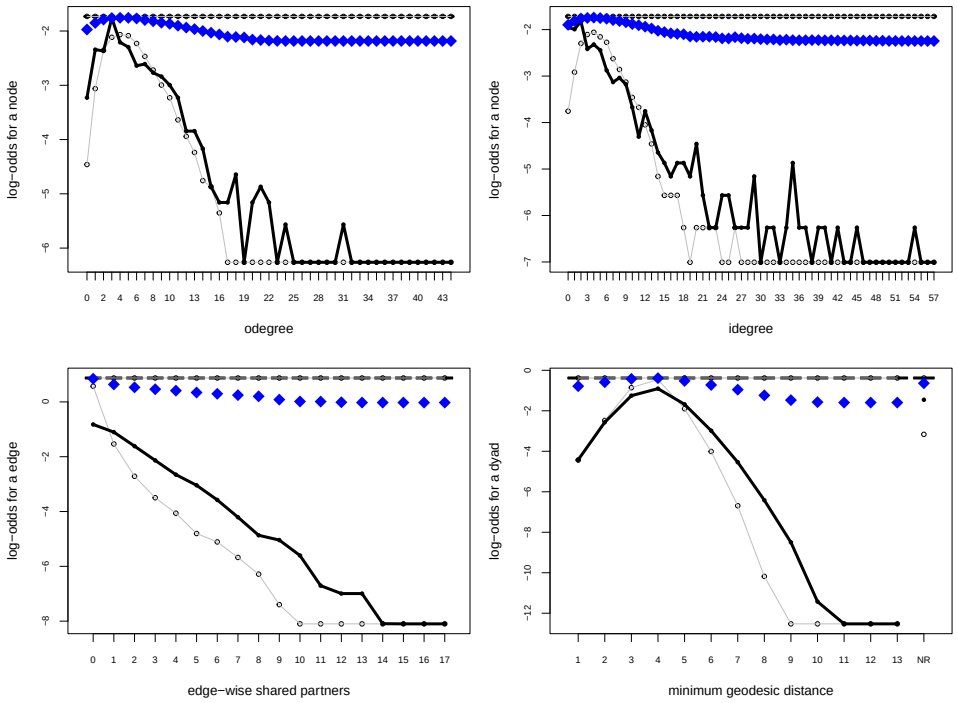


Figura 3.1: Grafici di alcune statistiche di rete per la rete osservata di nati nel XVIII secolo e quelle simulate dal modello stimato. I *boxplot* racchiudono il 95% delle statistiche simulate, i diamanti blu sono la loro media, le linee a punti rappresentano i relativi intervalli di confidenza e la linea nera raffigura le statistiche osservate nella rete originale.

quindi come la statistica *twopath* fornisca una valutazione delle possibili triangolazioni, mentre *gwesp* sia effettivamente un conteggio pesato di tali interazioni fra tre nodi. Il modello presenta dunque sia statistiche markoviane sia quantità che catturano una dipendenza locale più estesa (Tabella 3.2). Dal valore dei relativi parametri stimati, sembra emergere una bassa propensione a creare connessioni indirette, ma un'elevata tendenza a chiudere i pochi potenziali triangoli che si vengono a formare. Inoltre, la rete non è molto densa, ma c'è una buona propensione alla reciprocità. Le figure umanistiche e istituzionali tendono a connettersi in maniera maggiore rispetto agli artisti, mentre per gli altri domini questa tendenza non è significativamente differente. Si evidenzia una spiccata omofilia tra nodi che condividono dominio o secolo di nascita, mentre uomini e donne non presentano differenze rilevanti nella connettività. L'osservazione è resa però inconcludente dalla presenza di tre sole donne nella rete.

La verifica grafica della bontà di adattamento non offre risultati molto diversi da quelli ottenuti nella sezione precedente: lo studio di simulazione mostra distribuzioni molto concentrate, con la media aritmetica, distante dalla mediana, esterna ai *boxplot*. Anche qui, le statistiche della rete osservata si trovano spesso a cavallo dell'intervallo di confidenza, la distribuzione delle statistiche simulate appare anomala e il comportamento degli stimatori e delle relative varianze sembra distorto. Per questo motivo, anche dove le quantità osservate cadono all'interno degli intervalli, si ritiene di avere sufficienti elementi per affermare che l'adattamento del modello è, se

	Stima	Std. Error	Pr(>|z|)
edges	-5.53	0.73	0.00
reciprocity	2.27	0.38	0.00
twopath	-0.20	0.08	0.01
gwesp	0.52	0.18	0.00
U_{domain}	1.59	0.24	0.00
$U_{century}$	1.41	0.27	0.00
$domain_{Humanities}$	0.36	0.15	0.02
$domain_{Institutions}$	0.31	0.15	0.03
$domain_{Science}$	-0.24	0.55	0.67
$domain_{Sports}$	-0.15	0.17	0.40
$gender_{Male}$	0.34	0.32	0.30

Tabella 3.2: Valore dei coefficienti, con relativi *standard error* e *p-value*, del modello stimato per i personaggi irlandesi

non insufficiente, sicuramente troppo instabile.

Ciò si rivela anche in un confronto grafico sommario tra la rete osservata e alcune simulate dal modello stimato (Figura 3.3): le reti simulate non sembrano del tutto diverse da quella osservata, ma presentano alcune differenze. Difatti, la presenza di alcuni nodi isolati è bilanciata da altri nodi particolarmente connessi: entrambe sono caratteristiche assenti nella rete originale. Si può ipotizzare che le reti simulate seguano la legge di potenza in modo più fedele di quella osservata. La differenza non sta quindi tanto nel numero di archi, o nella loro reciprocità, quanto nella distribuzione degli stessi tra i vari nodi in maniera più elitaria rispetto alla rete originale.

3.3.4 Una rete selezionata pseudo-casualmente

Un'ultima rete presa in esame è stata selezionata pseudo-casualmente dalla rete Pantheon. Il termine "pseudo-casuale" si riferisce al fatto che la selezione dei nodi è avvenuta in tre stadi: la prima fase di selezione dei nodi dal *dataset* è stata puramente casuale; a causa dell'ordine della rete e della sua sparsità, ciò avrebbe comportato una sotto-rete estremamente sparsa, se non vuota. Al primo stadio è quindi seguita una seconda fase in cui sono stati selezionati tutti i nodi connessi ai primi, per garantire la presenza di archi nella sottorete; infine, si è proceduto con una sorta di potatura (*pruning*) della sottorete, per rimuovere i nodi dal grado troppo ridotto.

La rete finale è costituita da 111 attori, che rispecchiano in modo sorprendentemente fedele le caratteristiche principali rilevate sulla rete Pantheon. C'è una preponderanza di attori nati negli ultimi due secoli e provenienti dal mondo occidentale, ma la rappresentazione di altre aree geografiche o periodi storici non è assente. Le *views* ricevute dalle biografie dei soggetti selezionati seguono un andamento esponenziale, e le donne sono presenti in quantità anche superiore rispetto a quanto lo siano nella rete intera. Infine, la presenza di artisti è leggermente superiore a

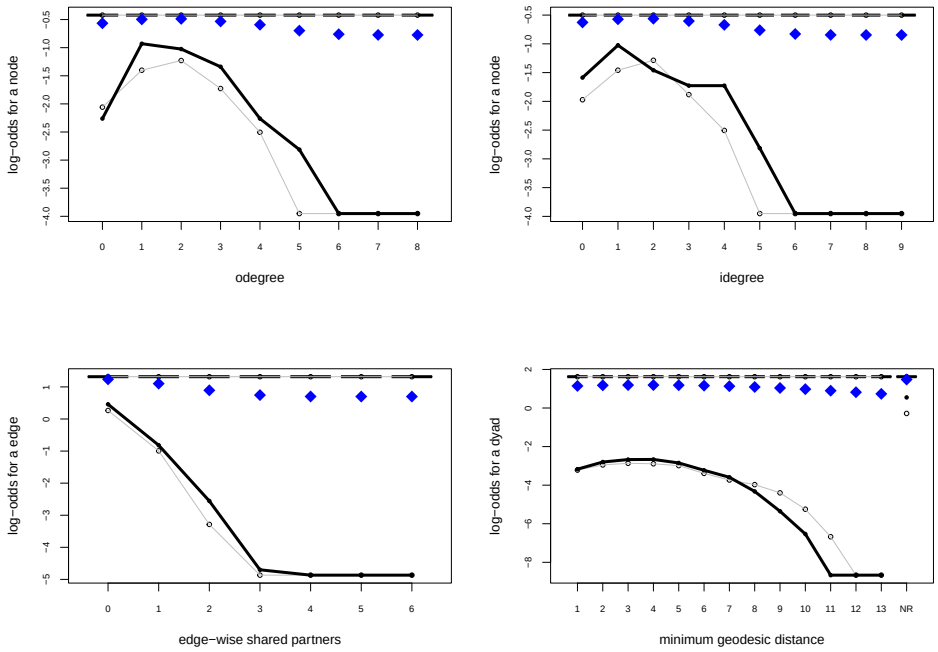


Figura 3.2: Grafici di alcune statistiche di rete per la rete osservata di personaggi irlandesi e quelle simulate dal modello stimato. I *boxplot* racchiudono il 95% delle statistiche simulate, i diamanti blu sono la loro media, le linee a punti rappresentano i relativi intervalli di confidenza e la linea nera raffigura le statistiche osservate nella rete originale.

quanto atteso, ma gli altri domini sono rappresentati in proporzioni fedeli alla rete completa.

La rete risulta poco densa anche in questo caso, con una reciprocità molto pronunciata (Tabella 3.3); ciò è in parte dovuto anche al metodo di selezione. La tendenza a creare connessioni indirette, o *twopath*, non sembra elevata né significativa, così come la distribuzione dell'*in-degree* pesata geometricamente (*geometrically weighted in-degree distribution*, o $gwID$). Nonostante non sia significativa, essa viene tenuta nel modello per ragioni di stabilità. La distribuzione dell'*out-degree* pesata geometricamente (*geometrically weighted out-degree distribution*, o $gwOD$) è invece significativa e indica una rete più democratica, senza veri e propri *hub*. Per entrambe le statistiche pesate geometricamente si è usato un τ pari a 0.25, poiché forniva l'adattamento migliore. Il numero di *views* della pagina Wikipedia di una persona non sembra influenzare significativamente le sue connessioni, al contrario della condivisione degli attributi di nodo: attori accomunati dagli stessi dominio, genere, continente o secolo di nascita hanno infatti una probabilità di connessione indubbiamente più elevata. In questa sottorete, le figure del mondo umanistico, istituzionale e pubblico creano una quantità di connessioni significativamente superiore agli artisti; scienziati ed esploratori hanno una connettività a essi sostanzialmente analoga, mentre gli sportivi si connettono di meno, anche se, pure qui, non vi è significatività statistica.

L'analisi dei grafici sulla bontà di adattamento non sembra mostrare differenze dai due casi visti in precedenza: una gran parte delle distribuzioni delle statistiche nelle reti simulate varia poco, risultando in *boxplot* schiacciati, ma alcune reti *outlier*

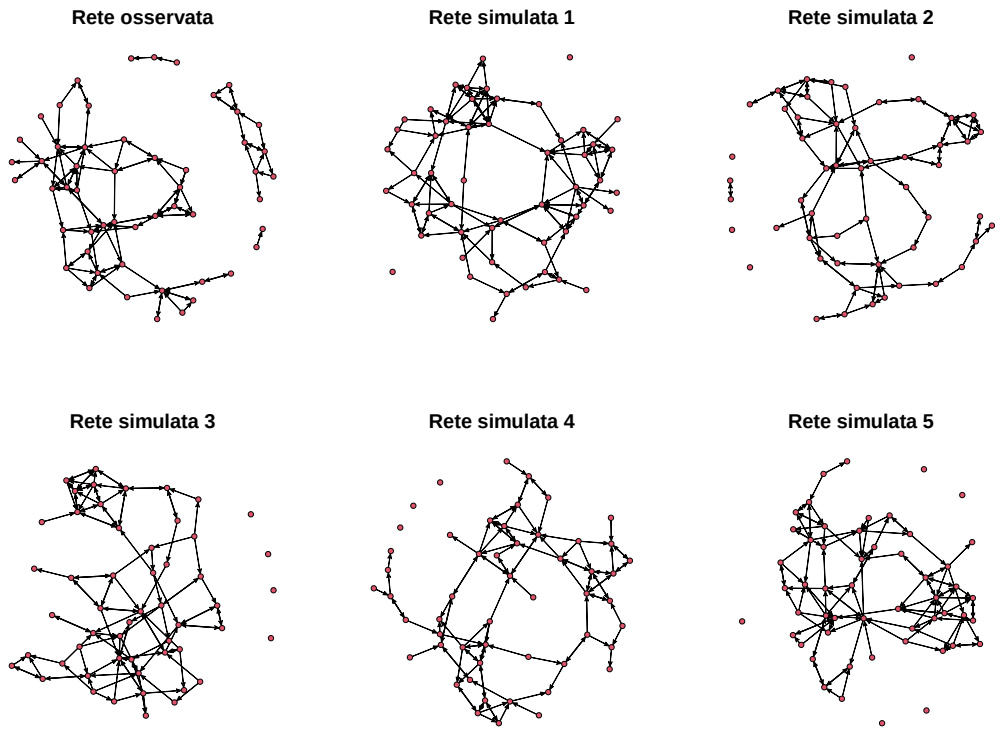


Figura 3.3: La rete osservata dei personaggi irlandesi e 5 reti simulate dal modello stimato.

portano la media delle quantità osservate a essere più bassa ed estremamente variabile (Figura 3.4). Le statistiche nella rete osservata si rivelano essere nuovamente lungo gli estremi degli intervalli di confidenza; sembra ragionevole concludere anche in questo caso che il modello implichi una distribuzione di reti piuttosto anomala, e, di conseguenza, fondamentalmente inaffidabile nella descrizione del meccanismo generatore della rete osservata, o di reti a essa simili.

Un confronto grafico tra la rete osservata e alcune simulate dal modello stimato evidenzia che quest'ultimo sembra cogliere in maniera soddisfacente solo alcune tendenze della rete originale (Figura 3.5). Densità e reciprocità sembrano essere riprodotte in modo abbastanza fedele, e la saltuaria presenza di nodi isolati non appare inconciliabile con i dati osservati. Nelle reti simulate vi sono però zone ad alta densità, che non trovano un corrispettivo nella rete originale; essa appare, come già menzionato, più equa e democratica nella distribuzione degli archi. Lo squilibrio nella distribuzione del grado dei nodi delle reti simulate è da ritenersi troppo sistematico per sostenere che il modello stimato sia adeguato.

Emerge da questa sezione una certà difficoltà nella stima sia di ERGM stabili e non degeneri, che di modelli con un adattamento ai dati soddisfacente e consistente. Gli ERGM appaiono flessibili, ma la scelta delle statistiche da includere nei modelli non è sempre intuitiva, così come la loro interpretazione, soprattutto quantitativa.

	Stima	**Std. Error**	**Pr(>|z|)**
$edges$	-7.35	0.39	0.00
$reciprocity$	4.60	0.27	0.00
$twopath$	0.07	0.04	0.14
$gwID$	0.03	0.40	0.95
$gwOD$	2.21	0.61	0.00
$views$	$2 \cdot 10^{-7}$	$2 \cdot 10^{-7}$	0.31
U_{domain}	1.65	0.15	0.00
$U_{continent}$	0.28	0.12	0.02
$U_{century}$	0.75	0.15	0.00
U_{gender}	0.40	0.15	0.01
$domain_{Exploration}$	0.25	0.42	0.55
$domain_{Humanities}$	0.52	0.13	0.00
$domain_{Institutions}$	0.25	0.10	0.01
$domain_{PublicFigure}$	0.78	0.30	0.01
$domain_{Science}$	0.09	0.18	0.62
$domain_{Sports}$	-0.20	0.16	0.20

Tabella 3.3: Valore dei coefficienti, con relativi *standard error* e *p-value*, del modello stimato per la rete selezionata pseudo-casualmente

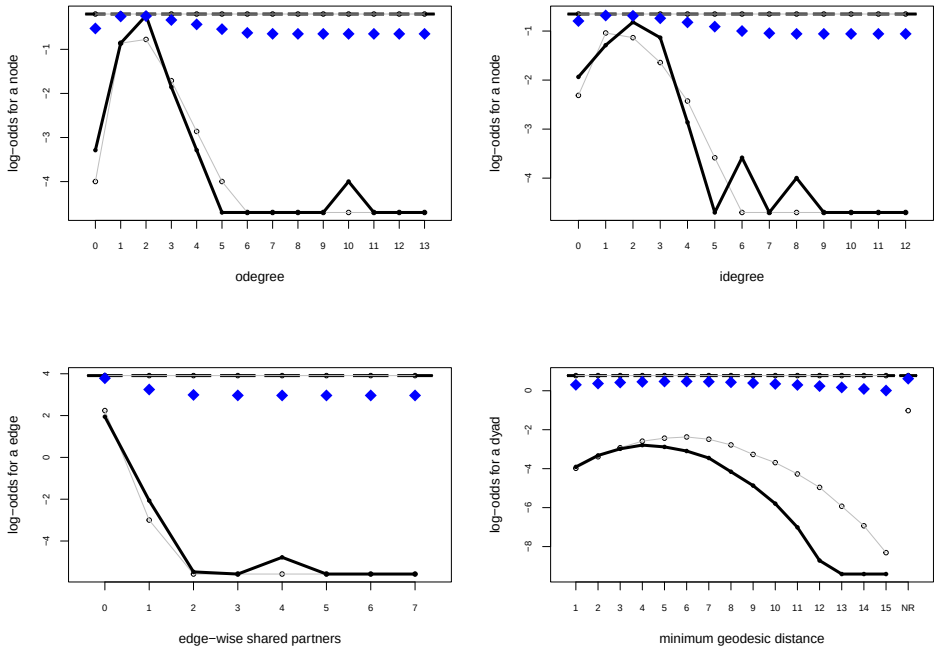


Figura 3.4: Grafici di alcune statistiche di rete per la rete osservata selezionata pseudo-casualmente e quelle simulate dal modello stimato. I *boxplot* racchiudono il 95% delle statistiche simulate, i diamanti blu sono la loro media, le linee a punti rappresentano i relativi intervalli di confidenza e la linea nera raffigura le statistiche osservate nella rete originale.

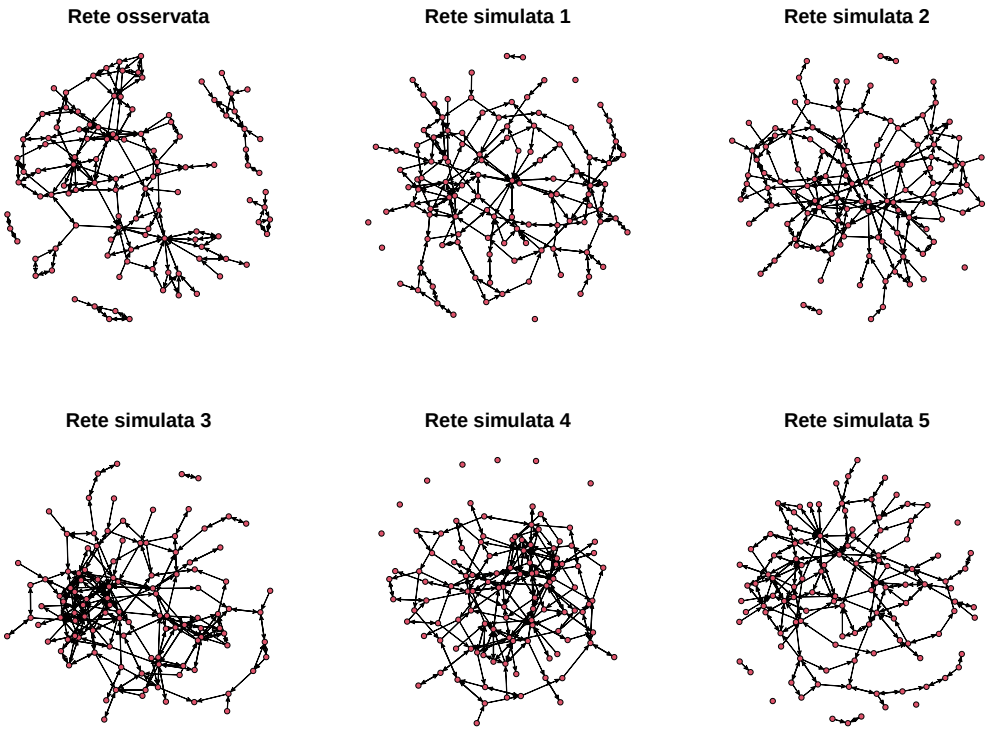


Figura 3.5: La rete selezionata pseudo-casualmente e 5 reti simulate dal modello stimato.

Conclusioni

In questo elaborato si è svolta un'analisi dei dati Pantheon e dei *pattern* che sono emersi al loro interno, sia nello spazio che nel tempo. Sebbene la composizione del *dataset* possa suggerire una prospettiva occidentalocentrica, una valutazione obiettiva in questo senso ricade più nell'ambito delle scienze storico-sociali che in quello delle scienze statistiche. Si è osservato come il numero grezzo di *views* di una biografia Wikipedia non sia un buon indicatore dell'importanza storica del relativo personaggio; al contrario, l'indice HPI sembra offrire delle *performance* più affidabili.

Sono poi state presentate le reti statistiche, illustrandone alcune proprietà e quantità notevoli. Si sono introdotte delle matrici per misurare, in termini assoluti e relativi, le interazioni tra *cluster* interni alle reti. Questi strumenti hanno permesso un'esplorazione approfondita della rete Pantheon, partendo da una prospettiva globale per poi analizzare le sottoreti e i loro legami. Il confronto tra diversi indici di rilevanza ha evidenziato delle relazioni tra di essi; alcuni indici sembrano portare a conclusioni ragionevoli e non sempre distorte a favore del mondo occidentale.

È stata affrontata la problematica della visualizzazione delle reti, offrendo una panoramica sulle soluzioni proposte in letteratura. L'applicazione di algoritmi di *layout* sulla rete Pantheon e sulle sue sottoreti ha consentito di osservare i comportamenti sia dei nodi che dei gruppi indotti da alcune variabili. I risultati grafici ottenuti appaiono coerenti con le indagini condotte.

Le analisi evidenziano una spiccata omofilia, che interessa principalmente i domini d'appartenenza dei soggetti nel *dataset*. Ovviamente anche la condivisione del periodo storico o dell'area geografica si sono dimostrati fattori influenti nella formazione di legami. Le osservazioni riguardo le differenze tra i generi sono inconcludenti e probabilmente risentono del *selection bias* nella composizione dei dati.

Nell'ultimo capitolo sono stati presentati alcuni modelli per i dati di rete, con una particolare attenzione dedicata agli ERGM. Questa classe di modelli, pur presentandosi come una soluzione elegante e flessibile, ha rivelato una serie di problematiche a cui i modelli stimati su dei gruppi di nodi di Pantheon non sono stati immuni. La *degeneracy*, la scelta delle statistiche di rete da includere nel modello e la complessa interpretazione dei parametri sono solo alcune delle difficoltà riscontrate. I modelli stimati hanno mostrato un comportamento incostante e inaffidabile, che mette in dubbio la possibilità di trarre conclusioni con un buon grado di fiducia.

Appare evidente la necessità di trovare strade alternative. Si potrebbero provare altre combinazioni di statistiche nella formulazione degli ERGM, ma, soprattutto per le reti dirette e di grandi dimensioni, sembra che la ricerca sia ancora relativamente immatura. Sfruttare alcuni dei metodi di stima proposti più di recente è una possibilità, ma queste soluzioni vanno ancora testate in modo estensivo. Un'ulteriore

opzione è rappresentata dalla gamma di modelli alternativi che sono stati menzionati: data la struttura apprezzabilmente ripartita della rete Pantheon, un modello a blocchi, ed in particolare un *mixed membership stochastic block model*, può risultare adeguato. Occorre tuttavia considerare la notevole dimensione della rete, che rappresenta una sfida significativa per gli algoritmi attuali. Questa complessità ostacola la stima di modelli abbastanza sofisticati per cogliere le dinamiche del processo generatore della rete stessa.

Di ulteriore interesse risulta l'analisi del *dataset* Pantheon 2.0, in attesa che un lavoro analogo a quello di Beytía e Schobin (2018) venga svolto per esplorare la rete aggiornata. I nuovi dati sono infatti decisamente più abbondanti, ed è possibile che coprano meglio alcune aree geografiche o alcuni periodi storici, offrendo una prospettiva meno distorta. Anche alcune domande sulle differenze di genere, più che mai attuali, potrebbero trovare risposte più sicure nel *dataset* più recente.

Appendice A

Grafici aggiuntivi

In questa appendice sono riportate le *heatmap* delle matrici P, Q e T per alcune sottoreti analizzate nell'elaborato (si vedano le formule 2.4 e 2.5).

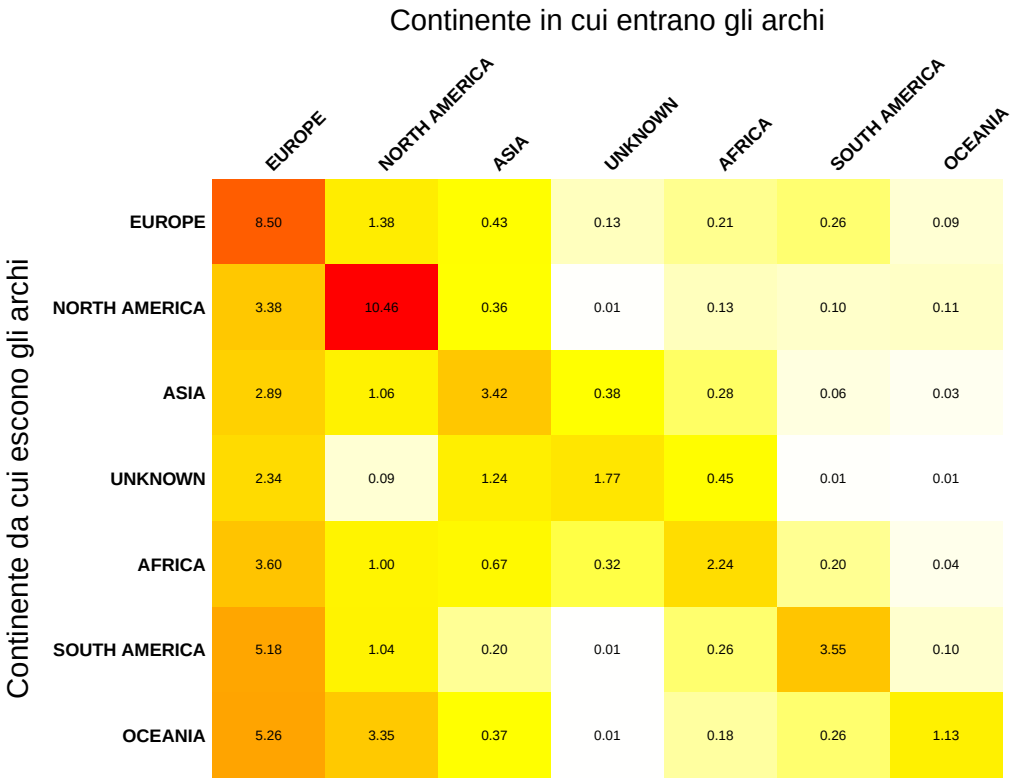


Figura A.1: Matrice P per i continenti

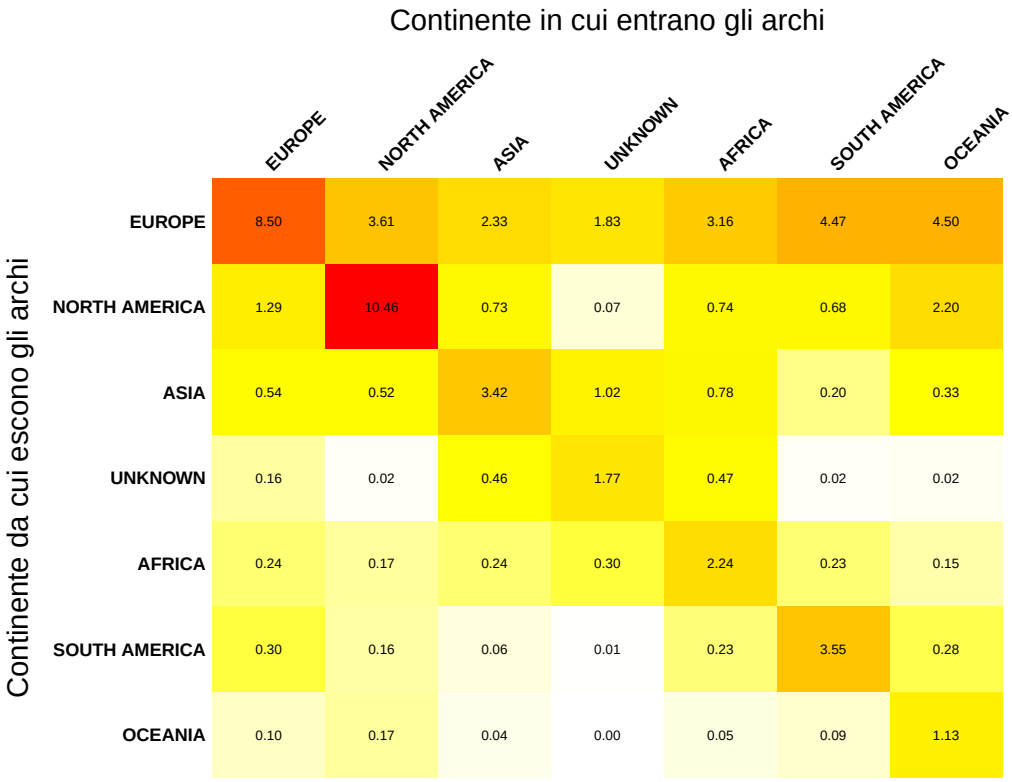


Figura A.2: Matrice Q per i continenti

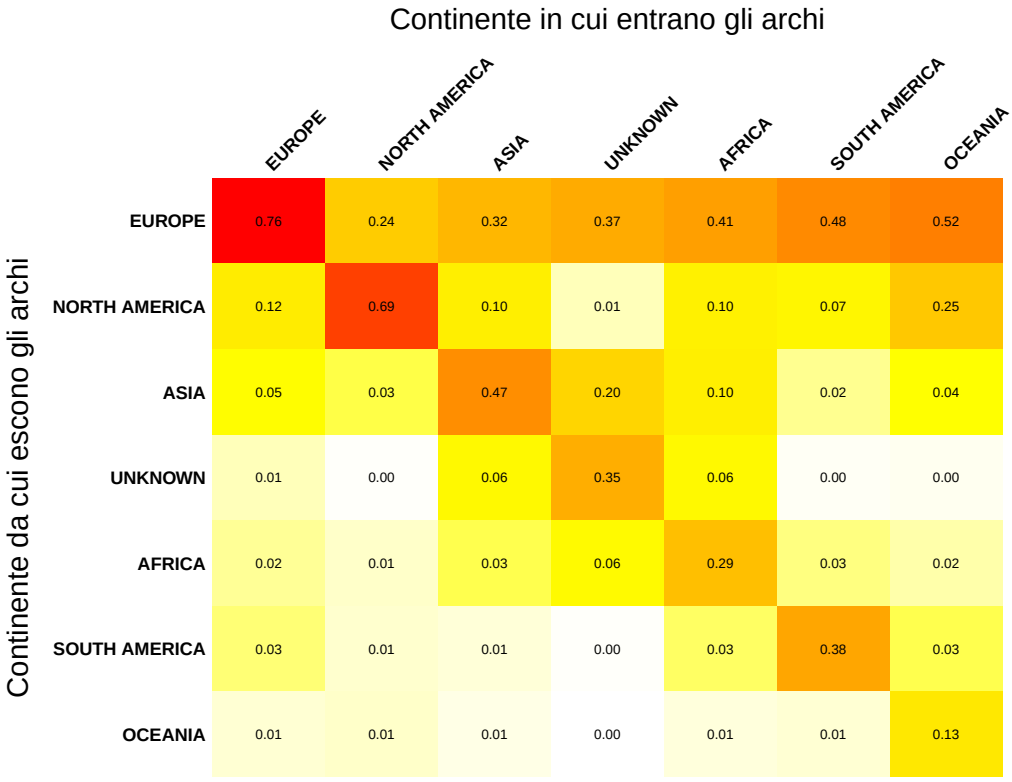


Figura A.3: Matrice T per i continenti

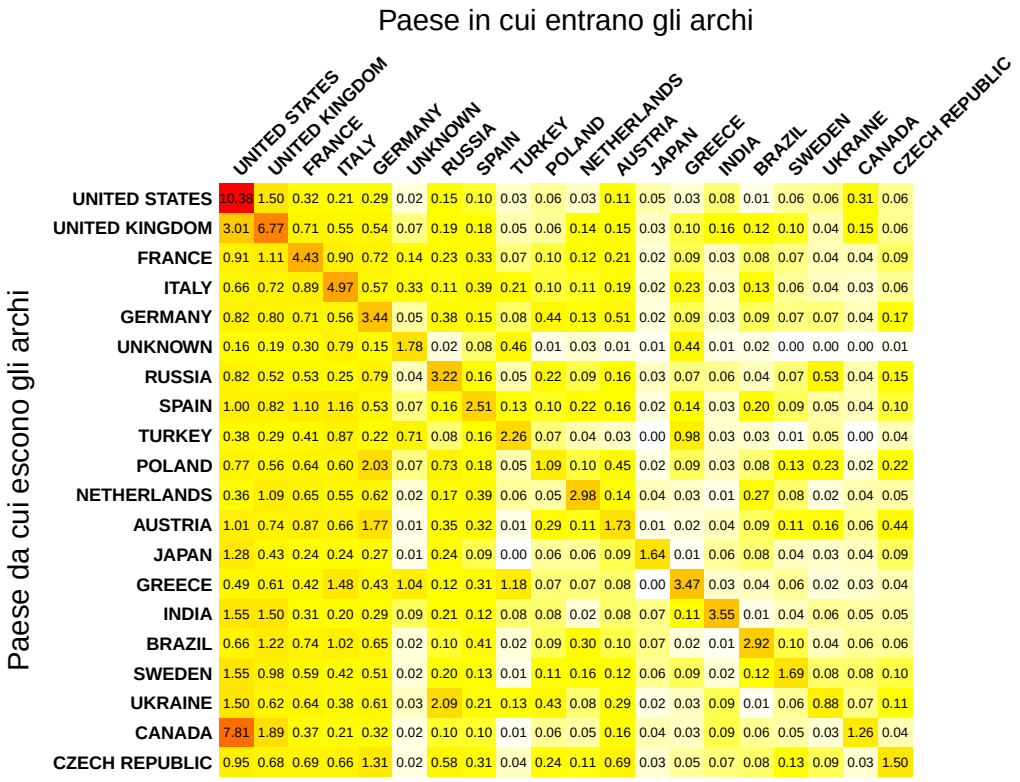


Figura A.4: Matrice P per i Paesi

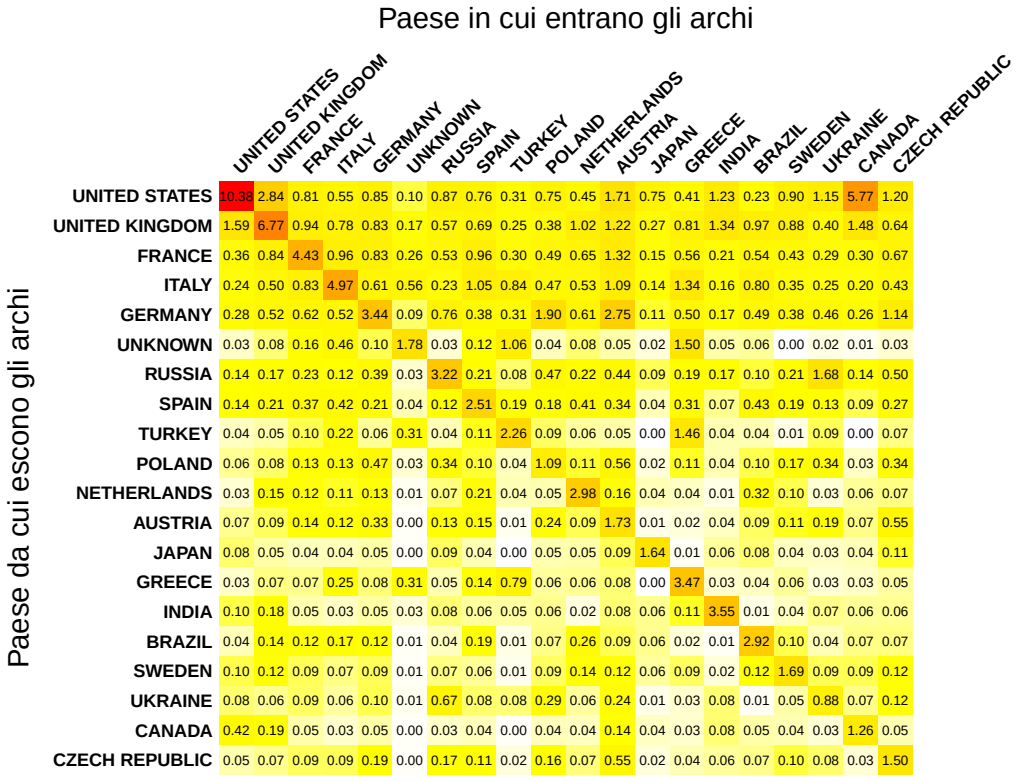


Figura A.5: Matrice Q per i Paesi

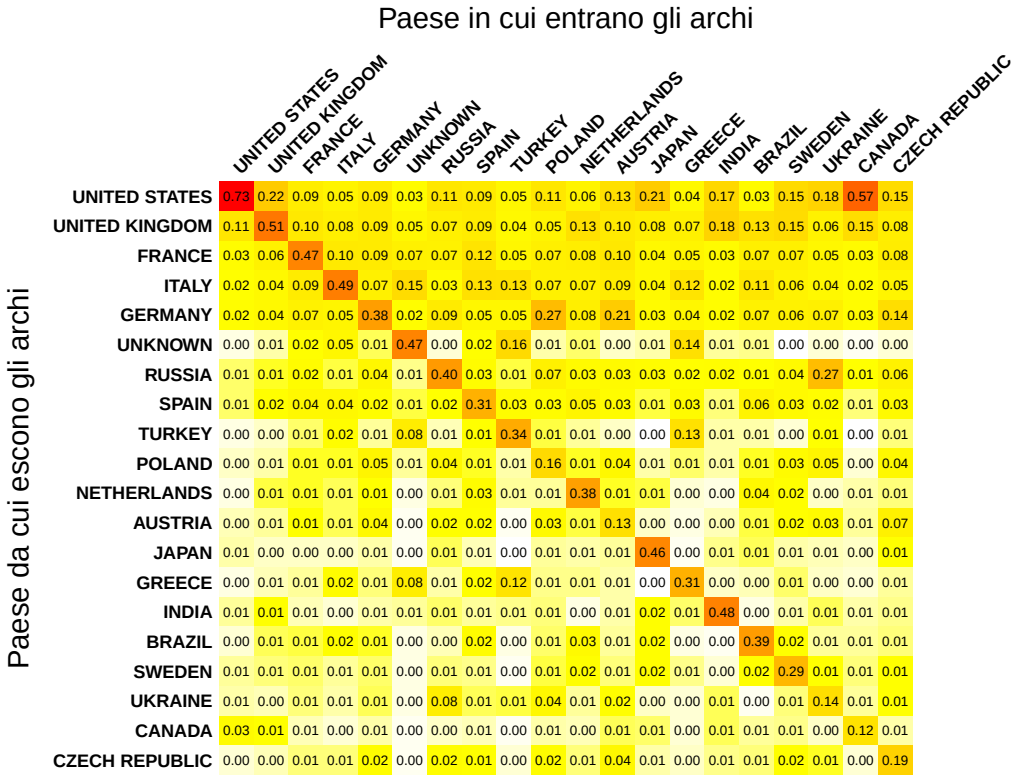


Figura A.6: Matrice T per i Paesi

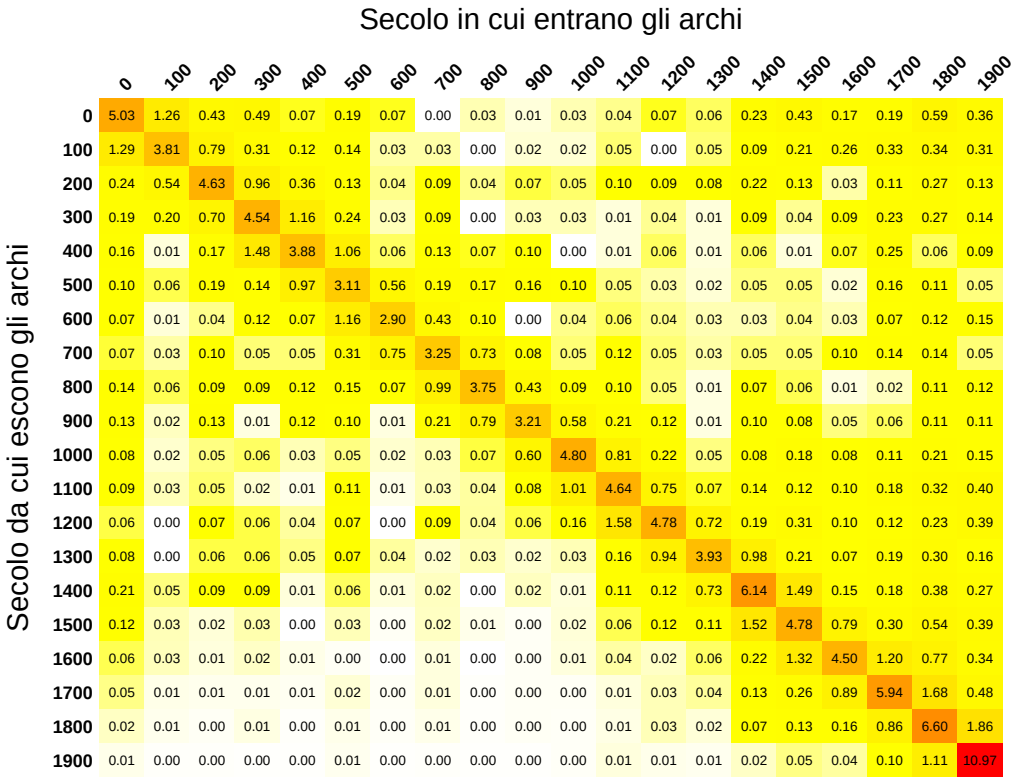


Figura A.7: Matrice P per i secoli

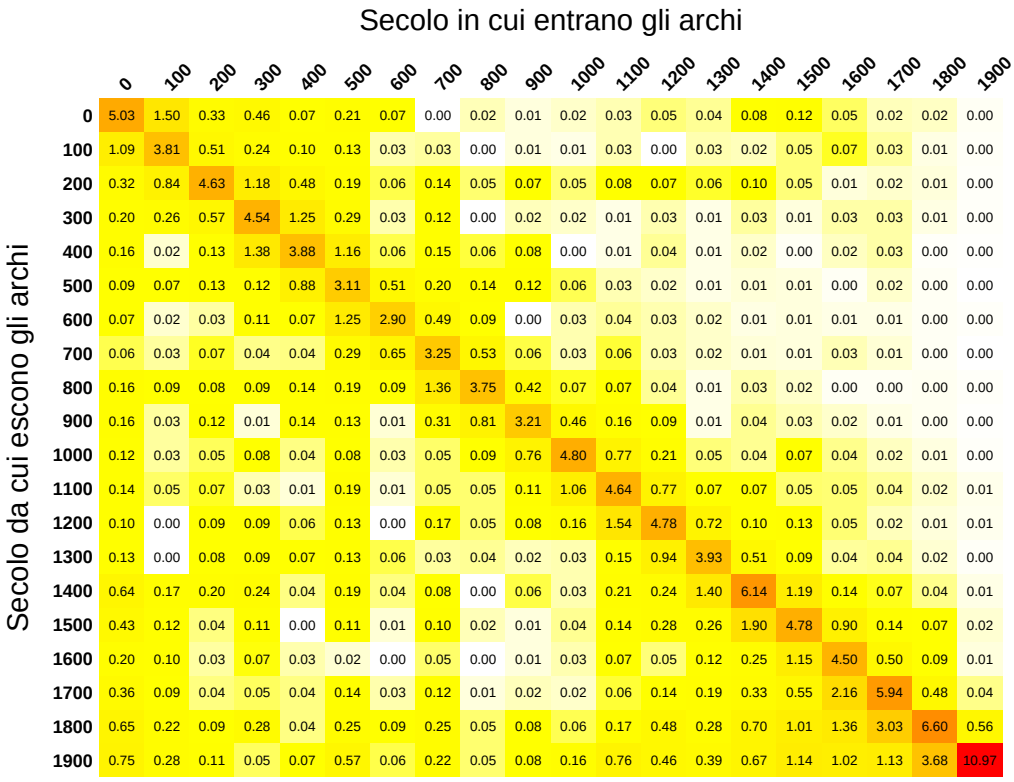


Figura A.8: Matrice Q per i secoli

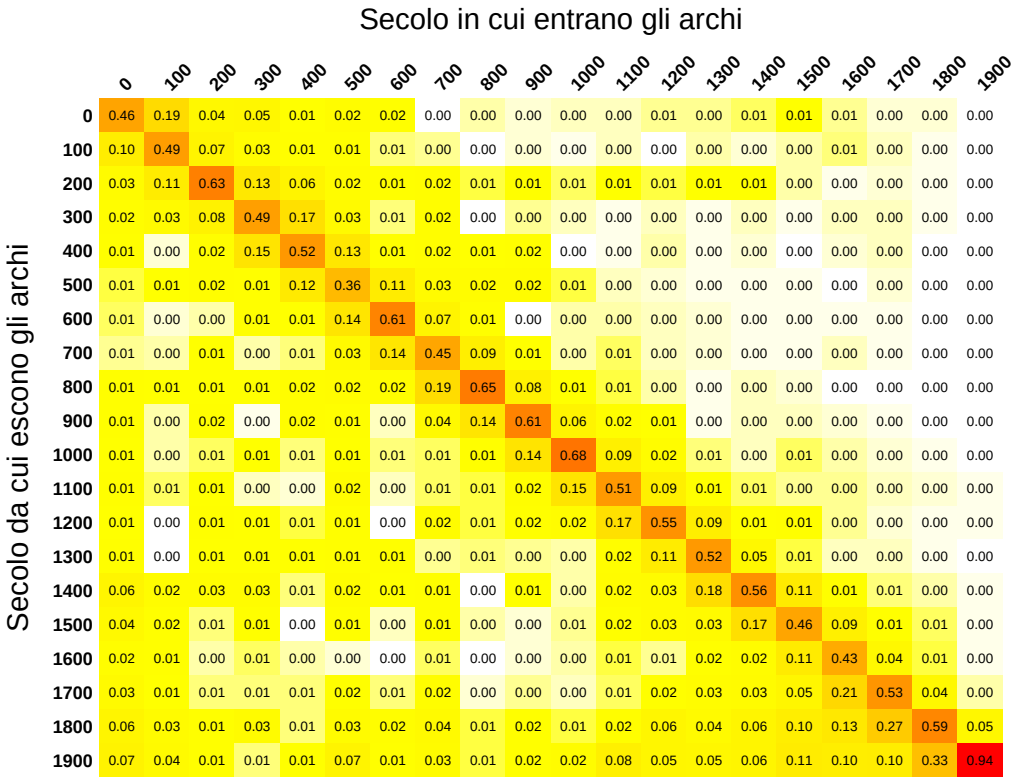


Figura A.9: Matrice T per i secoli

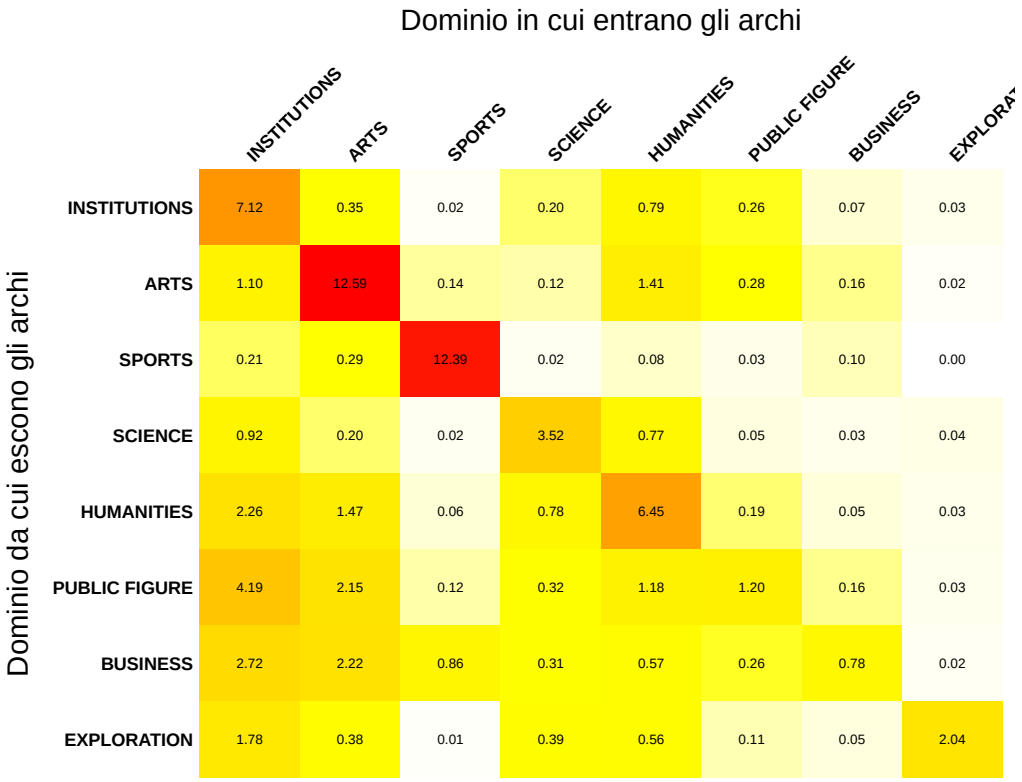


Figura A.10: Matrice P per i domini

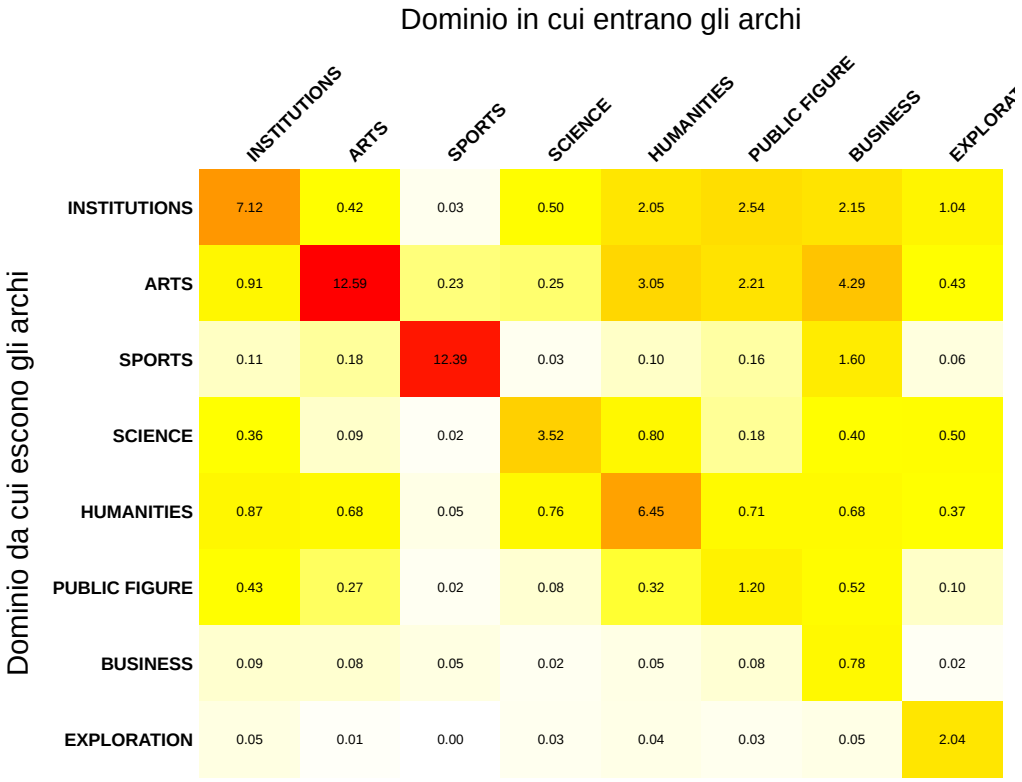


Figura A.11: Matrice Q per i domini

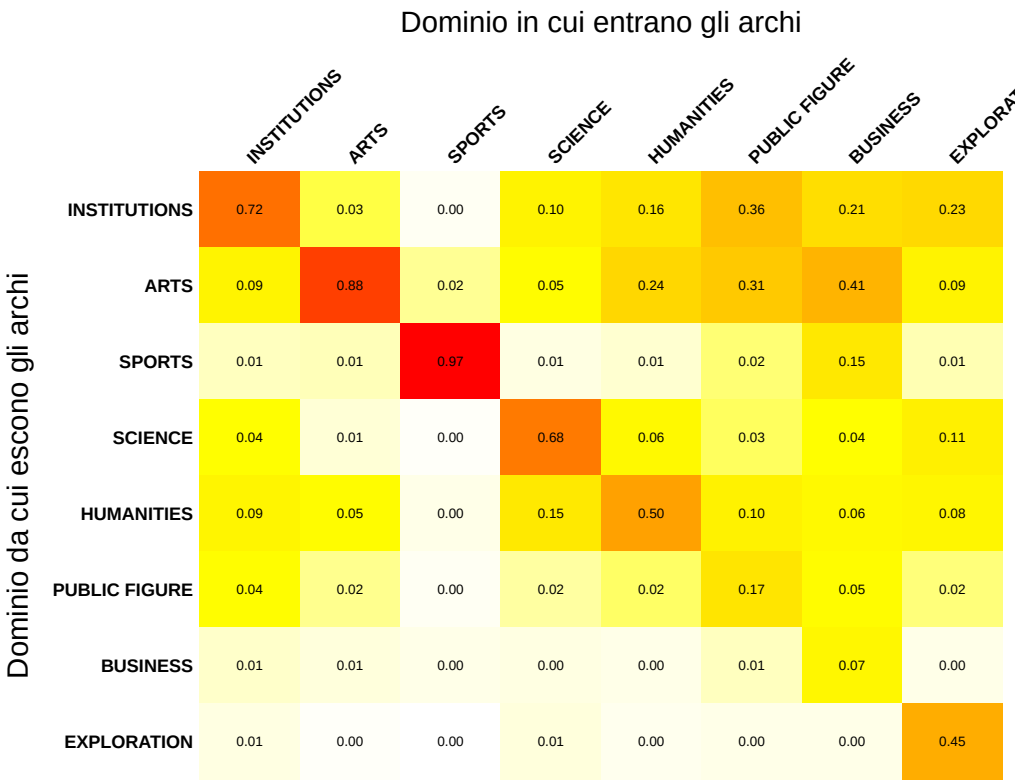


Figura A.12: Matrice T per i domini

Bibliografia

Barabási A.-L.; Albert R. (1999). Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.

Bastian M.; Heymann S.; Jacomy M. (2009). Gephi: An open source software for exploring and manipulating networks.

Beytía P.; Schobin J. (2018). Networked pantheon: A relational database of globally famous people. *SSRN Electronic Journal*.

Bonacich P. (1987). Power and centrality: A family of measures. *American journal of sociology*, **92**(5), 1170–1182.

Brin S.; Page L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, **30**(1-7), 107–117.

Brown A. (2011). Wikipedia as a data source for political scientists: Accuracy and completeness of coverage. *PS: Political Science & Politics*, **44**.

Butts C. T. (2008). network: a package for managing relational data in r. *Journal of Statistical Software*, **24**(2).

Butts C. T. (2015). *network: Classes for Relational Data*. The Statnet Project (`http://www.statnet.org`). R package version 1.13.0.1.

Csardi G.; Nepusz T. (2006). The igraph software package for complex network research. *InterJournal*, **Complex Systems**, 1695.

Csárdi G.; Nepusz T.; Traag V.; Horvát S.; Zanini F.; Noom D.; Müller K. (2024). *igraph: Network Analysis and Visualization in R*. R package version 2.0.3.9043.

Erdős P.; Rényi A. (1959). On random graphs i. *Publ. math. debrecen*, **6**(290-297), 18.

Frank O.; Strauss D. (1986). Markov graphs. *JASA. Journal of the American Statistical Association*, **81**.

Goodreau S. (2007). Advances in exponential random graph (p*) models applied to large social networks. *Social networks*, **29**, 231–248.

Goodreau S. M.; Handcock M. S.; Hunter D. R.; Butts C. T.; Morris M. (2008). A statnet tutorial. *Journal of Statistical Software*, **24**(9), 1–26.

Handcock M. (2003a). Assessing degeneracy in statistical models of social networks. *Journal of the American Statistical Association*, **76**, 33–50.

Handcock M. S. (2003b). Statistical models for social networks: Inference and degeneracy. In *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*. The National Academies Press.

Handcock M. S.; Hunter D. R.; Butts C. T.; Goodreau S. M.; Morris M. (2008). statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, **24**(1), 1–11.

Handcock M. S.; Hunter D. R.; Butts C. T.; Goodreau S. M.; Krivitsky P. N.; Morris M. (2023). *ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. The Statnet Project (`https://statnet.org`). R package version 4.6.0.

Hoff P. D.; Raftery A. E.; Handcock M. S. (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association*, **97**(460), 1090–1098.

Holland P. W.; Leinhardt S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the american Statistical association*, **76**(373), 33–50.

Holme P. (2019). Rare and everywhere: Perspectives on scale-free networks. *Nature communications*, **10**(1), 1016.

Hu Y. (2005). Efficient, high-quality force-directed graph drawing. *Mathematica journal*, **10**(1), 37–71.

Hunter D. R. (2007). Curved exponential family models for social networks. *Social networks*, **29**(2), 216–230.

Hunter D. R.; Handcock M. S.; Butts C. T.; Goodreau S. M.; Morris M. (2008a). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, **24**(3), 1–29.

Hunter D. R.; Goodreau S. M.; Handcock M. S. (2008b). Goodness of fit of social network models. *Journal of the american statistical association*, **103**(481), 248–258.

Jacomy M.; Venturini T.; Heymann S.; Bastian M. (2014). Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one*, **9**(6), e98679.

Kitchens B.; Johnson S.; Gray P. (2020). Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption. *MIS Quarterly*, **44**.

Kleinberg J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, **46**(5), 604–632.

Krivitsky P. N.; Hunter D. R.; Morris M.; Klumb C. (2023). ergm 4: New features for analyzing exponential-family random graph models. *Journal of Statistical Software*, **105**(6), 1–44.

Lovekar K.; Sengupta S.; Paul S. (2021). Testing for the network small-world property.

Mcpherson M.; Smith-Lovin L.; Cook J. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, **27**, 415–.

Milgram S. (1967). The small world problem. *Psychology today*, **2**(1), 60–67.

Moreno J. L.; Jennings H. H. (1938). Statistics of social configurations. *Sociometry*, **1**(3/4), 342–374.

Pattison P.; Robins G. (2002). Neighborhood–based models for social networks. *Sociological methodology*, **32**(1), 301–337.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rochat Y. (2009). Closeness centrality extended to unconnected graphs: The harmonic centrality index. In *Asna*.

Salter-Townshend M.; White A.; Gollini I.; Murphy T. B. (2012). Review of statistical network analysis: models, algorithms, and software. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **5**(4), 243–264.

van Der Pol J. (2019). Introduction to network modeling using exponential random graph models (ergm): theory and an application using r-project. *Computational Economics*, **54**(3), 845–875.

Van Duijn M. A.; Snijders T. A.; Zijlstra B. J. (2004). p2: a random effects model with covariates for directed graphs. *Statistica Neerlandica*, **58**(2), 234–254.

Wasserman S.; Faust K. (1994). *Social Network Analysis: Methods and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press.

Watts D. J.; Strogatz S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, **393**(6684), 440–442.

Wickham H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Yu A.; Ronen S.; Hu K.; Lu T.; Hidalgo C. (2016). Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific data*, **3**, 150075.

Ringraziamenti

Non posso concludere questo lavoro senza esprimere la mia gratitudine alle persone che mi sono state vicine in questo percorso.

Il primo grazie va ai miei genitori. Grazie per avermi supportato in tutto e per tutto, per aver creduto in me e per avermi permesso di vivere questa bellissima esperienza accademica e di vita, anche se forse non avete ancora capito cosa studio. Spero di rendervi orgogliosi ogni giorno, anche solo in minima parte rispetto a quanto io lo sono di avervi con me.

Un grandissimo ringraziamento va alla nonna Sandra, che sento sempre qui con me e a cui questo traguardo è dedicato. La sua importanza nel mio percorso di crescita non è esprimibile a parole, e la sua assenza in questo momento si fa sentire un po' di più.

Un ringraziamento speciale anche ai miei nonni, Giulia e Nello. Siete sempre presenti per me e mi fate sentire il vostro amore incondizionato in ogni momento. Spero che assistere a questo momento speciale possa ripagarvi di tutto quello che avete fatto e fate per me.

Ringrazio di cuore Patrizia e Giorgio. Siete un luogo sicuro dove posso aprirmi ed essere me stesso, e mi offrite la spensieratezza di cui ogni tanto ho bisogno.

Un grazie collettivo alla mia famiglia, ai prozii, agli zii ed ai cugini. Vi ringrazio per la vostra presenza e l'interesse che mostrate costantemente per il mio percorso. Sappiate che non lo do per scontato.

Un grande grazie va alle mie coinquiline, Camilla, Marilyn e Matilde. Siete diventate una seconda famiglia, e vi sono grato per avermi sopportato anche quando non sembrava possibile. Siete la nota più dolce della mia vita padovana, e avete reso via Novara una Casa.

Voglio poi ringraziare tutti gli amici che ci sono stati per me, anche a distanza: Emanuele, Alice, Benedetta, Francesca, Sabrina, Silvia, Silvia, Silvio, Thea. Avervi accanto significa molto per me, la vostra amicizia è sincera e preziosa. So di poter contare su di voi.

Un grazie sentito a tutti i miei amici del Dipartimento. Siete troppi da elencare, ma siete stati dei compagni di viaggio incredibili. Anche se le nostre strade accademiche si separano, spero che le nostre vite non facciano altrettanto e mantengano una correlazione significativa.

Infine, ringrazio Davide e Iacopo per avermi permesso di vivere questo giorno speciale in compagnia di Amici veri. Auguri a noi tre, la componente principale di questo traguardo.