

# Analisi della rete dei blog politici USA (2004)

Gabriele Fontanot

2025-07-19

## 1. Introduzione e motivazione

L'analisi delle reti sociali online è fondamentale per comprendere i meccanismi di formazione delle opinioni, la diffusione delle informazioni e la polarizzazione. In questo lavoro, attraverso il caso studio dei blog politici USA, siamo andati ad indagare se la struttura delle connessioni rispecchi la divisione ideologica della società digitale dell'epoca. La blogosfera politica americana del 2004 rappresenta, quindi, un esempio emblematico di ecosistema sociale online, caratterizzato da fenomeni di polarizzazione.

### Domande di ricerca:

- La struttura delle connessioni riflette la polarizzazione politica?
  - Chi sono i blog più influenti o centrali nella rete?
  - Esistono blog “ponte” tra le fazioni?
  - Le community individuate coincidono con gli schieramenti politici?
  - La rete mostra proprietà globali tipiche delle reti sociali reali?
- 

## 2. Descrizione del dataset

Il dataset utilizzato è, come già anticipato, la rete dei blog politici statunitensi raccolti durante la campagna elettorale del 2004 (fonte: Adamic & Glance, 2005).

I nodi rappresentano dei blog, ciascuno etichettato con l'orientamento politico (liberal = 0, conservative = 1); gli archi invece rappresentano i link tra i vari blog, con una rete diretta.

```
library(igraph)
library(ggraph)
library(ggplot2)

g <- read_graph("C:/Users/gabri/OneDrive/Desktop/Università/Magistrale/Primo anno/Advanced Data science
summary(g)

## IGRAPH be30289 D--- 1490 19090 --
## + attr: id (v/n), label (v/c), value (v/n), source (v/c)
```

```

cat("Numero nodi:", vcount(g), "\n")

## Numero nodi: 1490

cat("Numero archi:", ecount(g), "\n")

## Numero archi: 19090

table(V(g)$value)

##
##      0      1
## 758 732

```

## 2.1 Statistiche di base

Analizziamo le statistiche di base e la struttura generale della rete:

```

# Statistiche base
nodi <- vcount(g)
archi <- ecount(g)
comp <- components(g)
num_componenti <- comp$no
dim_componente_principale <- max(comp$csizes)
nodi_isolati <- sum(degree(g, mode = "all") == 0)
cat("Numero componenti connesse:", num_componenti, "\n")

## Numero componenti connesse: 268

cat("Dimensione componente principale:", dim_componente_principale, "\n")

## Dimensione componente principale: 1222

cat("Numero nodi isolati:", nodi_isolati, "\n")

## Numero nodi isolati: 266

```

Dalle informazioni di base emerge la presenza di una componente principale molto grande, che suggerisce l'esistenza di un core della blogosfera fortemente interconnesso. Il numero elevato di nodi isolati riflette invece la poca importanza di molti blog rispetto al dibattito centrale.

## 2.2 Visualizzazione della struttura

Applichiamo un primo layout per osservare la struttura generale e la divisione politica. Abbiamo effettuato anche un focus sulla componente principale, il core della nostra rete.

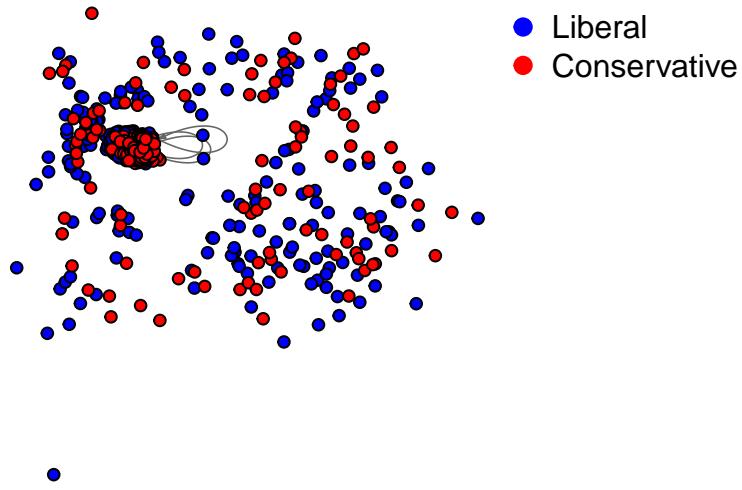
```

set.seed(123)
plot(
  g,
  layout = layout_with_fr,
  vertex.label = NA,
  vertex.size = 5,
  vertex.color = ifelse(V(g)$value == 0, "blue", "red"),
  edge.color = "grey40",           # archi
  edge.width = 0.8,               # spessore archi
  edge.arrow.size = 0.2,          # grandezza freccia
  main = "Rete Intera"
)

legend(
  "topright", legend = c("Liberal", "Conservative"),
  col = c("blue", "red"), pch = 19, pt.cex = 1.2, bty = "n"
)

```

## Rete Intera



```

# Componente principale
# comp già calcolata
giant <- induced_subgraph(g, which(comp$membership == which.max(comp$csize)))

set.seed(123)
plot(
  giant,

```

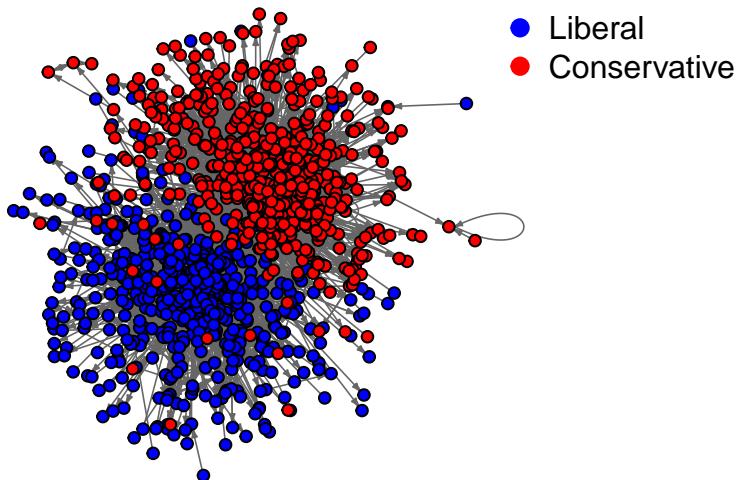
```

layout = layout_with_fr,
vertex.label = NA,
vertex.size = 5,
vertex.color = ifelse(V(giant)$value == 0, "blue", "red"),
edge.color = "grey40",
edge.width = 0.8,
edge.arrow.size = 0.2,
main = "Componente principale"
)

legend(
  "topright", legend = c("Liberal", "Conservative"),
  col = c("blue", "red"), pch = 19, pt.cex = 1.2, bty = "n"
)

```

## Componente principale



La visualizzazione evidenzia una netta separazione tra la componente principale e una grande quantità di nodi periferici, confermando quanto osservato nelle statistiche di base. I colori dei nodi mettono subito in risalto la distribuzione degli orientamenti politici, mostrando la presenza di tendenze di aggregazione. Visualizziamo anche la presenza di due grandi aree corrispondenti agli orientamenti politici, ma non una separazione assoluta: sono visibili anche collegamenti trasversali tra i gruppi. Questo suggerisce una polarizzazione significativa, ma non totale, della blogosfera.

### 3. Analisi locale

Vogliamo identificare i blog più influenti attraverso diverse metriche di centralità: in-degree (popolarità), betweenness (nodi ponte) e PageRank (autorevolezza globale).

#### 3.1 Risultati principali

```
#Analisi delle metriche di centralità

deg_in <- degree(g, mode = "in")
betw <- betweenness(g, directed = TRUE)
pager <- page_rank(g, directed = TRUE)$vector

#ordino i nodi e scelgo gli indici
top_in_indices <- order(deg_in, decreasing = TRUE)[1:10]
top_betw_indices <- order(betw, decreasing = TRUE)[1:10]
top_pager_indices <- order(pager, decreasing = TRUE)[1:10]

#creo i dataframe per le top10 delle misure
top_in <- data.frame(
  label = V(g)$label[top_in_indices],
  in_degree = deg_in[top_in_indices],
  orientamento = V(g)$value[top_in_indices]
)
top_betw <- data.frame(
  label = V(g)$label[top_betw_indices],
  betweenness = betw[top_betw_indices],
  orientamento = V(g)$value[top_betw_indices]
)
top_pager <- data.frame(
  label = V(g)$label[top_pager_indices],
  pagerank = pager[top_pager_indices],
  orientamento = V(g)$value[top_pager_indices]
)

print(top_in)

##          label in_degree orientamento
## 1      dailykos.com        338          0
## 2    instapundit.com       277          1
## 3 talkingpointsmemo.com     269          0
## 4   atrios.blogspot.com     264          0
## 5   drudgereport.com       240          1
## 6 powerlineblog.com        221          1
## 7 blogsforbush.com        212          1
## 8 washingtonmonthly.com     201          0
## 9 michellemalkin.com       201          1
## 10 truthlaidbear.com       187          1

print(top_betw)
```

```

##          label betweenness orientamento
## 1      blogsforbush.com  218480.74      1
## 2      atrios.blogspot.com  91011.59      0
## 3      instapundit.com   76177.68      1
## 4      dailykos.com     54819.21      0
## 5 newleftblogs.blogspot.com  45886.95      0
## 6 madkane.com/notable.html  44996.54      0
## 7      wizbangblog.com   40631.62      1
## 8      lashawnbarber.com  36076.33      1
## 9      hughhewitt.com    34222.60      1
## 10     washingtonmonthly.com 32671.60      0

```

```
print(top_pager)
```

```

##          label pagerank orientamento
## 1      dailykos.com 0.017897495      0
## 2      atrios.blogspot.com 0.015189152      0
## 3      instapundit.com 0.012593268      1
## 4      blogsforbush.com 0.012460222      1
## 5 talkingpointsmemo.com 0.012402045      0
## 6 michellemalkin.com 0.010882831      1
## 7 drudgereport.com 0.010684616      1
## 8 washingtonmonthly.com 0.010518799      0
## 9 powerlineblog.com 0.008912599      1
## 10 andrewsullivan.com 0.008591861      1

```

Visualizziamo i risultati anche con un grafico a barre per i top 10 blog per le tre misure, particolarizzati sempre per orientamento politico.

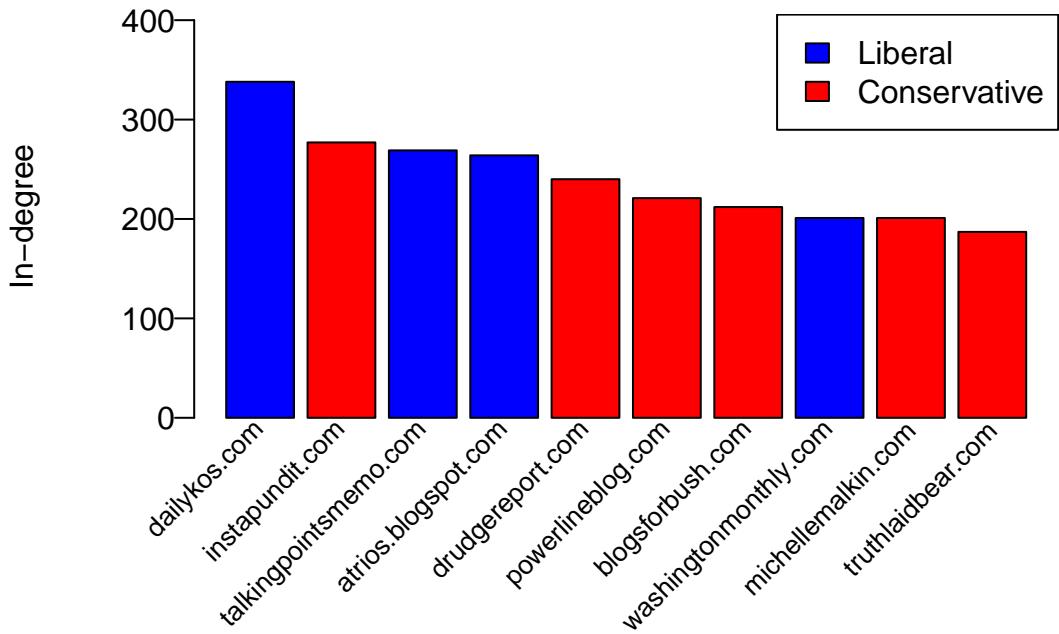
```

par(mar = c(8, 8, 4, 2))
par(mgp = c(4, 0.5, 0))

bp_in <- barplot(
  top_in$in_degree,
  names.arg = rep("", length(top_in$label)),
  las = 2,
  cex.names = 0.8,
  col = ifelse(top_in$orientamento == 0, "blue", "red"),
  main = "Top 10 Blog per In-degree",
  ylab = "In-degree",
  ylim = c(0, max(top_in$in_degree) * 1.2)
)
text(
  x = bp_in, y = par("usr")[3] - 0.02 * diff(par("usr")[3:4]),
  labels = top_in$label, srt = 45, adj = 1, xpd = TRUE, cex = 0.8
)
legend("topright", legend = c("Liberal", "Conservative"), fill = c("blue", "red"))

```

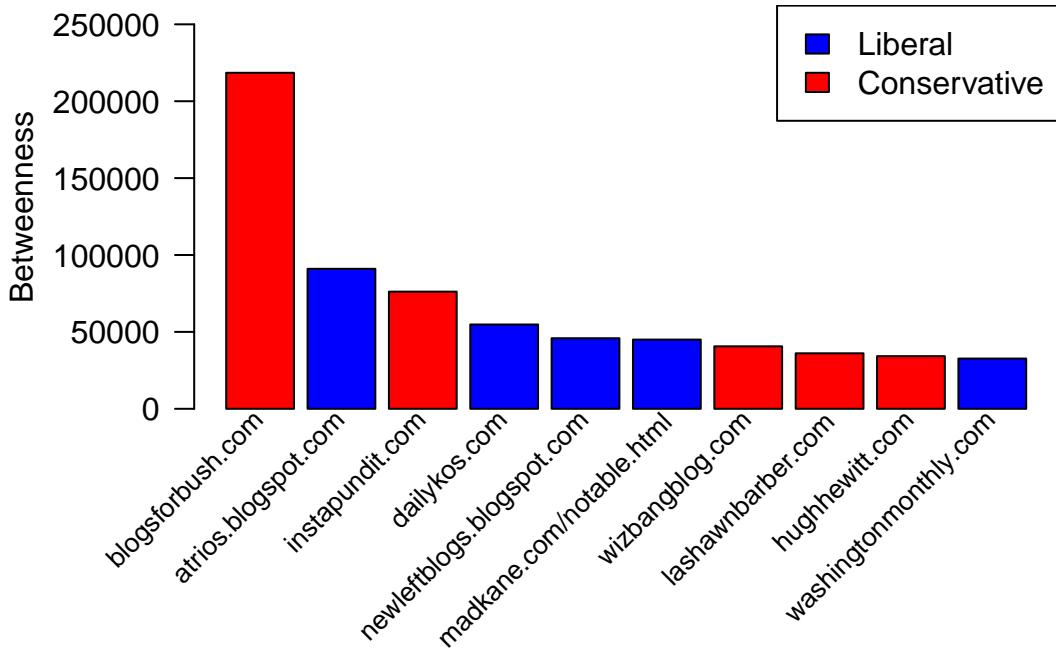
## Top 10 Blog per In-degree



```
par(mar = c(8, 8, 4, 2)) # (bottom, left, top, right)
par(mgp = c(4, 0.9, 0))

bp_betw <- barplot(
  top_betw$betweenness,
  names.arg = rep("", length(top_betw$label)),
  las = 2,
  cex.names = 0.8,
  col = ifelse(top_betw$orientamento == 0, "blue", "red"),
  main = "Top 10 Blog per Betweenness",
  ylab = "Betweenness",
  ylim = c(0, max(top_betw$betweenness) * 1.2)
)
text(
  x = bp_betw, y = par("usr")[3] - 0.02 * diff(par("usr"))[3:4],
  labels = top_betw$label, srt = 45, adj = 1, xpd = TRUE, cex = 0.8
)
legend("topright", legend = c("Liberal", "Conservative"), fill = c("blue", "red"))
```

## Top 10 Blog per Betweenness



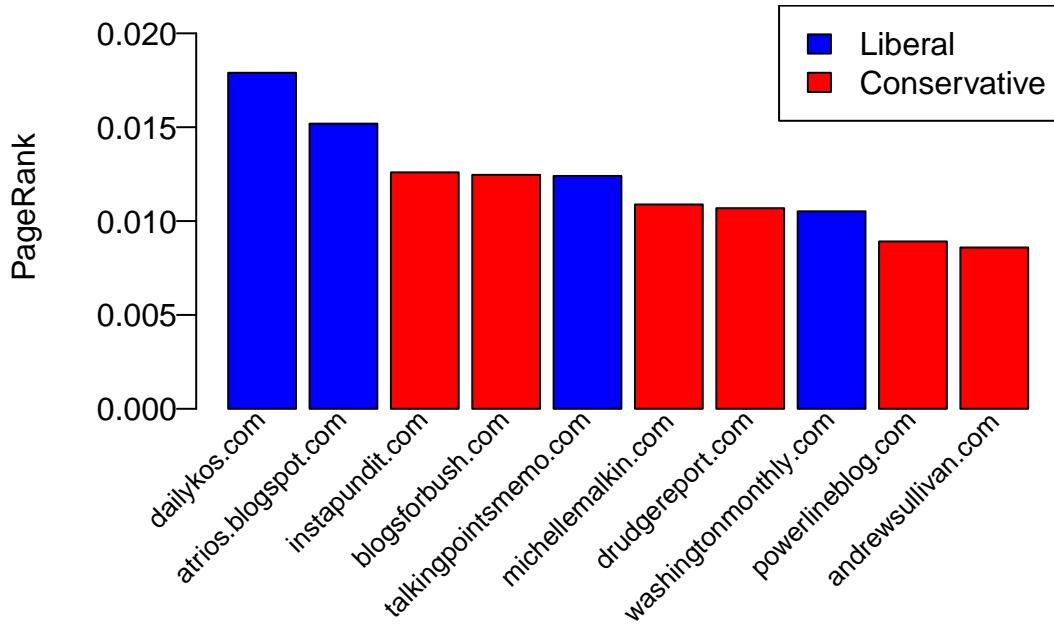
```

par(mar = c(8, 8, 4, 2)) # (bottom, left, top, right)
par(mgp = c(4, 0.5, 0)) #per etichetta

bp <- barplot(
  top_pager$pagerank,
  names.arg = rep("", length(top_pager$label)),
  las = 2,
  cex.names = 0.8,
  col = ifelse(top_pager$orientamento == 0, "blue", "red"),
  main = "Top 10 Blog per PageRank",
  ylab = "PageRank",
  ylim = c(0, max(top_pager$pagerank) * 1.2)
)
text(
  x = bp, y = par("usr")[3] - 0.02 * diff(par("usr"))[3:4]),
  labels = top_pager$label, srt = 45, adj = 1, xpd = TRUE, cex = 0.8
)
legend("topright", legend = c("Liberal", "Conservative"), fill = c("blue", "red"))

```

## Top 10 Blog per PageRank



- I blog con in-degree più alto sono quelli maggiormente citati, veri e propri centri di attenzione nelle varie conversazioni della blogosfera.

- I blog con betweenness più elevata svolgono invece il ruolo di “ponti” tra diverse aree della rete: pur non essendo sempre i più popolari, sono cruciali per la circolazione delle informazioni e il collegamento tra schieramenti diversi.
- Il PageRank permette di individuare i blog con maggiore autorevolezza complessiva, tenendo conto sia della quantità che della qualità delle citazioni ricevute.

Osserviamo che i ruoli di centralità non sono concentrati su un solo schieramento: entrambe le fazioni presentano blog chiave, e non sempre i più citati sono anche i più influenti nella rete nel suo complesso.

---

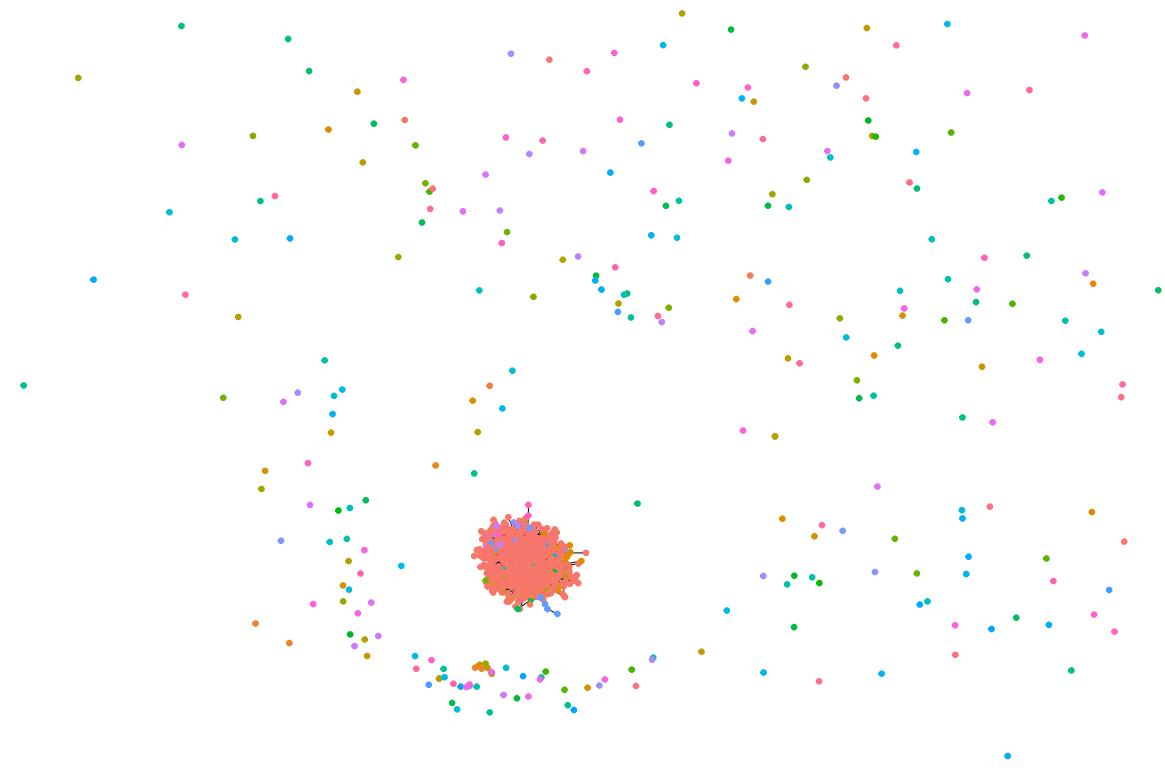
## 4. Community detection

L'algoritmo di clustering Infomap rileva le comunità ottimizzando la compressione delle traiettorie casuali sulla rete, ed è adatto per reti dirette e ampiamente usato per identificare community in contesti di informazione e flusso. Applicandolo sull'intera rete, la suddivisione in community appare frammentata, con numerose micro-componenti dovute ai molti nodi isolati o periferici e difficile da comprendere. Il passaggio successivo è stato infatti applicarlo alla componente principale, per riuscire ad estrarre informazioni più chiare.

## 4.1 Community detection sull'intera rete

```
comm_infomap <- cluster_infomap(g)
V(g)$community <- as.factor(membership(comm_infomap))
ggraph(g, layout = "fr") +
  geom_edge_link(alpha = 1, width = 0.05) +
  geom_node_point(aes(color = community), size = 0.5) +
  theme_void() +
  ggtitle("Community Detection (Infomap) - Intera Rete") +
  theme(legend.position = "none")
```

Community Detection (Infomap) – Intera Rete



## 4.2 Community detection sulla componente principale

Visualizziamo la struttura delle community sulla sola componente principale, per evitare l'influenza grafica dei nodi isolati e rendere il plot leggibile.

```
# comp e giant già calcolate
# Infomap su giant
set.seed(123)
comm_giant <- cluster_infomap(giant)
V(giant)$community <- as.character(membership(comm_giant))

tab <- sort(table(V(giant)$community), decreasing = TRUE)
```

```

top_clusters <- names(tab)[1:4]

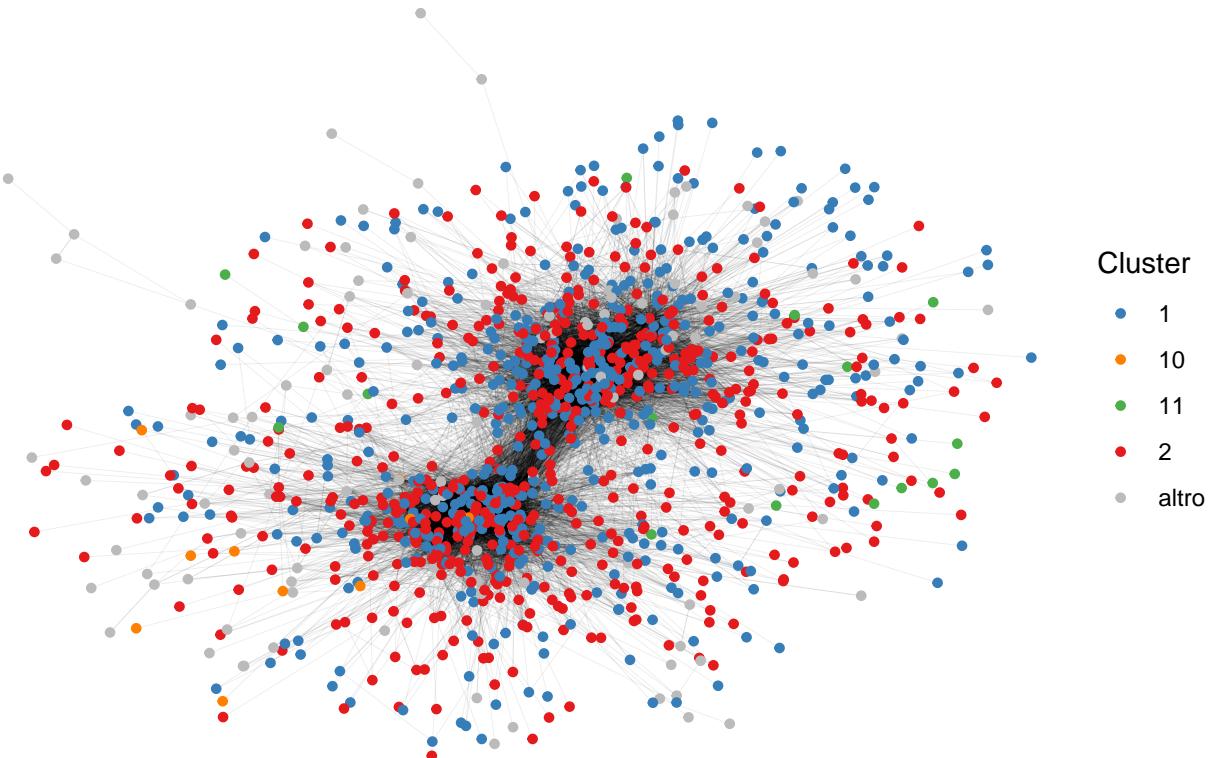
V(giant)$comm_main <- ifelse(V(giant)$community %in% top_clusters,
                                V(giant)$community, "altro")

palette <- c(
  setNames(
    c("#E41A1C", "#377EB8", "#4DAF4A", "#FF7F00"),
    top_clusters
  ),
  "altro" = "#BBBBBB"
)

ggraph(giant, layout = "fr") +
  geom_edge_link(alpha = 0.07, width = 0.08) +
  geom_node_point(aes(color = comm_main), size = 1.2) +
  scale_color_manual(values = palette, name = "Cluster") +
  theme_void() +
  ggtitle(sprintf("Cluster principali della componente gigante",
                  paste(top_clusters, collapse = ", "))) +
  theme(legend.position = "right")

```

## Cluster principali della componente gigante



```
cat("Distribuzione politica nei 4 cluster principali:\n")
```

```
## Distribuzione politica nei 4 cluster principali:
```

```

for (c in top_clusters) {
  idx <- which(V(giant)$community == c)
  num_lib <- sum(V(giant)$value[idx] == 0)
  num_con <- sum(V(giant)$value[idx] == 1)
  cat(sprintf("Cluster %s: Liberal = %d, Conservative = %d\n", c, num_lib, num_con))
}

## Cluster 2: Liberal = 277, Conservative = 276
## Cluster 1: Liberal = 225, Conservative = 304
## Cluster 11: Liberal = 12, Conservative = 4
## Cluster 10: Liberal = 10, Conservative = 0

```

Nella componente principale emergono cluster ben definiti, ma la composizione interna mostra una mescolanza tra orientamenti politici diversi. Ciò implica una polarizzazione strutturale solo parziale.

La differenza tra la visualizzazione sull'intera rete e quella sulla componente principale evidenzia quanto la presenza di nodi isolati e piccoli gruppi periferici possa andare a mascherare le vere community strutturali, cosa che non succede analizzando invece il core.

### 4.3 Incrocio community e orientamento politico

Incrociamo le community rilevate con l'orientamento politico per vedere se i gruppi corrispondono alle due divisioni ideologiche.

```

set.seed(123)
comm_giant <- cluster_infomap(giant)
V(giant)$community <- as.character(membership(comm_giant))

cross_tab_giant <- table(Community = V(giant)$community, Orientamento = V(giant)$value)
print(cross_tab_giant)

```

	Orientamento	
Community	0	1
1	225	304
10	10	0
11	12	4
12	4	1
13	3	1
14	4	0
15	2	0
16	2	0
17	2	0
18	2	2
19	3	0
2	277	276
20	3	0
21	2	0
22	2	0
23	1	2
24	2	0
25	2	0
26	3	1

```

##      27   0   3
##      28   0   4
##      29   0   6
##      3   8   0
##      30   0   3
##      31   0   2
##      32   0   2
##      33   0   2
##      34   0   2
##      35   0   5
##      36   0   5
##      37   0   3
##      38   0   2
##      39   0   2
##      4   5   1
##      40   0   2
##      5   3   0
##      6   3   0
##      7   1   1
##      8   3   0
##      9   2   0

```

Ci viene confermato che la suddivisione strutturale della rete non è allineata perfettamente agli schieramenti politici. Il valore di modularità basso rafforza questa osservazione, indicando che la polarizzazione è solo parziale e che esistono collegamenti significativi tra le due fazioni.

```

mod_val_giant <- modularity(comm_giant)
cat("Modularità della partizione (componente principale):", mod_val_giant, "\n")

```

---

```

## Modularità della partizione (componente principale): 0.02739014

```

## 5. Analisi globale

Calcoliamo le principali metriche sulla rete completa e sulla componente principale per confrontare l'intera rete e il nucleo sociale attivo.

Table 1: Confronto delle metriche globali: rete completa vs componente principale

Metrica	Rete.completa	Componente.principale
Numero nodi	1490.000	1222.000
Numero archi	19090.000	19089.000
Componenti connesse	268.000	1.000
Diametro	9.000	9.000
Lunghezza media del cammino	3.390	3.390
Reciprocità	0.240	0.240
Assortatività (politica)	0.823	0.823
Modularità (Infomap)	0.024	0.027

Le metriche globali risultano pressoché identiche tra l'intera rete e la sola componente principale. Questo è dovuto al fatto che la componente principale racchiude quasi tutta la struttura connessa e significativa della blogosfera, mentre il resto della rete è costituito da nodi isolati o piccoli gruppi che non contribuiscono alle distanze o alle proprietà strutturali di rilievo. Il confronto delle metriche mostra quindi che la quasi totalità delle proprietà strutturali dipende dalla componente principale, mentre i nodi periferici hanno impatto trascurabile.

## 5.1 Metriche

Le metriche globali della rete evidenziano una struttura tipica dei social network reali. Il diametro e la Lunghezza media del cammino sono bassi: ogni blog è raggiungibile in pochi passi, segno di una rete compatta e small world, che favorisce la rapida diffusione delle informazioni.

L'assortatività politica è alta: i blog preferiscono collegarsi ad altri con lo stesso orientamento, confermando la presenza di polarizzazione. Tuttavia, la modularità è bassa: le divisioni tra gruppi non sono così nette e permangono collegamenti tra aree diverse, segno di una polarizzazione non assoluta.

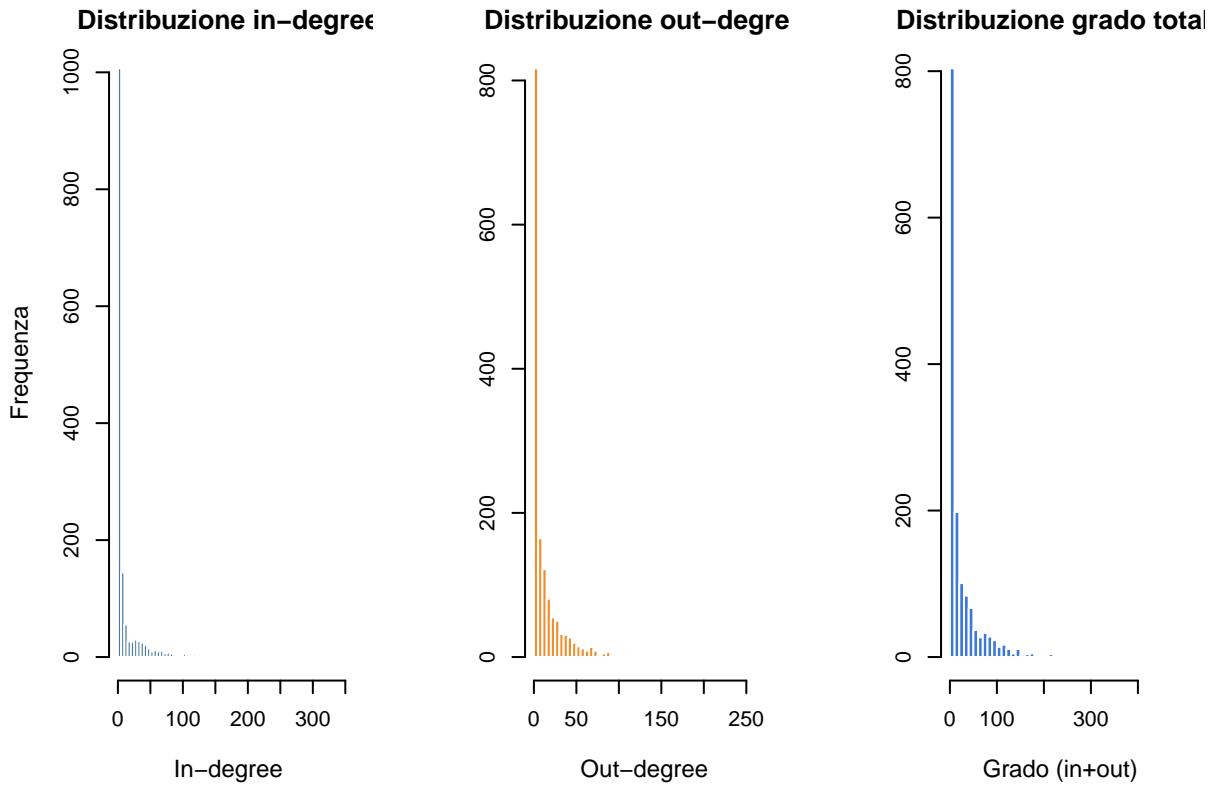
Infine, la reciprocità significativa indica che molte relazioni sono bidirezionali: nella blogosfera esistono non solo monologhi, ma anche dialoghi e scambi reciproci tra i diversi blog.

## 5.2 Proprietà di Small world e Scale-free

```
par(mfrow = c(1,3), mar = c(7,6,2,1))

deg_in <- degree(g, mode = "in")
deg_out <- degree(g, mode = "out")
deg_all <- degree(g, mode = "all")

hist(deg_in, breaks = 50, col = "#4e79a7", border = "white",
      main = "Distribuzione in-degree", xlab = "In-degree", ylab = "Frequenza",
      cex.main = 1.2, cex.lab = 1.1, cex.axis = 1)
hist(deg_out, breaks = 50, col = "#f28e2b", border = "white",
      main = "Distribuzione out-degree", xlab = "Out-degree", ylab = "",
      cex.main = 1.2, cex.lab = 1.1, cex.axis = 1)
hist(deg_all, breaks = 50, col = "#3C78D8", border = "white",
      main = "Distribuzione grado totale", xlab = "Grado (in+out)", ylab = "",
      cex.main = 1.2, cex.lab = 1.1, cex.axis = 1)
```



```
# Risetto il layout
par(mfrow=c(1,1))
```

- La maggior parte dei blog ha pochi collegamenti in entrata/uscita;
- Nodi con in-degree alto: blog molto citati, spesso “autorevoli” o di riferimento;
- Nodi con out-degree alto: blog molto attivi nel citare altri;
- Esistono alcuni super-hub (pochi blog con altissimo grado), ci sono pochissimi blog collegati a tanti altri e tantissimi blog con pochi collegamenti.
- **Small world:**
  - La distanza media tra i nodi (“Lunghezza media del cammino”) è molto bassa rispetto alla dimensione della rete.
  - Si può raggiungere ogni blog in pochi “passi” tramite i collegamenti, proprio come nei social network reali.
- **Scale-free:**
  - La distribuzione dei gradi mostra pochi nodi con tantissime connessioni (hub) e molti nodi con pochi collegamenti, seguendo una power law.
  - Quasi tutti i blog hanno pochi collegamenti, un numero minimo ne hanno moltissimi.
  - Questo favorisce la diffusione rapida delle idee, ma anche la concentrazione del potere informativo.

#### **Conseguenze:**

- La presenza di hub favorisce la diffusione rapida delle informazioni. - La struttura è robusta a rimozioni casuali ma fragile se si colpiscono gli hub principali.

La combinazione di proprietà ‘scale-free’ e ‘small-world’ rende la blogosfera un ambiente ideale per la rapida diffusione delle informazioni, ma la rende anche vulnerabile al controllo o alla rimozione dei nodi più centrali

---

## **6. Conclusioni**

L'analisi condotta permette di rispondere alle domande di ricerca iniziali, fornendo una visione articolata della blogosfera politica americana nel 2004. In sintesi:

- 1. La struttura delle connessioni riflette la polarizzazione politica?** Sì, ma non in modo assoluto: esistono due aree principali nella rete corrispondenti agli schieramenti liberal e conservative, tuttavia sono presenti anche collegamenti tra i due gruppi.
  - 2. Chi sono i blog più influenti o centrali nella rete?** I blog più influenti secondo in-degree, betweenness e PageRank sono distribuiti tra entrambi gli schieramenti. L'influenza è quindi un fenomeno trasversale e non limitato a una sola fazione.
  - 3. Esistono blog “ponte” tra le fazioni?** Sì, alcuni blog hanno valori elevati di betweenness, fungendo da ponti strategici che collegano le diverse community e facilitano lo scambio di informazioni tra aree ideologicamente distinte.
  - 4. Le community individuate coincidono con gli schieramenti politici?** Solo parzialmente: la rilevazione delle community mostra una mescolanza di blog di orientamenti diversi nei principali cluster, e la modularità bassa conferma che la polarizzazione non è netta a livello strutturale.
  - 5. La rete mostra proprietà globali tipiche delle reti sociali reali?** Sì, la rete presenta caratteristiche di piccolo mondo e scale-free, con una componente principale fortemente connessa che determina le proprietà globali della blogosfera.
- 

## **7. Limiti e sviluppi futuri**

- **Limiti dello studio:**

- L'analisi si basa su un'istantanea della blogosfera (2004) e non considera l'evoluzione temporale.
- Le community e i collegamenti sono studiati solo a livello strutturale, senza analizzare il contenuto effettivo dei blog.

- **Prospettive e sviluppi futuri:**

- Estendere l'analisi a una dimensione temporale diversa, osservando come cambia la struttura e la polarizzazione nel tempo.
  - Integrare informazioni sul contenuto dei blog (analisi semantica, sentiment analysis).
  - Applicare metodi di community detection più avanzati o comparare algoritmi diversi.
-

## Riferimenti

- Adamic, L. A., & Glance, N. (2005). *The political blogosphere and the 2004 US Election*. WWW-2005 Workshop on the Weblogging Ecosystem.
  - Dataset: <http://www-personal.umich.edu/~mejn/netdata/>
- 

Grazie per l'attenzione!