Network Analysis of Historical Fame

Exploring the Pantheon Dataset through Centrality, Homophily and ERGMs

Gabriele Paganelli • University of Padova

11 340	126 153	8	5 500+	3
Nodes	Directed Edges	Centrality Indices	Years Covered	ERGM Subnetworks

Overview

We apply statistical network analysis to the **Pantheon dataset**, a collection of 11 340 notable individuals drawn from Wikipedia (biography present in ≥25 languages). Directed edges between nodes are derived from hyperlinks between English-language biographies. The analysis spans exploratory description of node attributes, computation of centrality indices, visualization of the network, and inference via **Exponential Random Graph Models** on selected subnetworks.

Data

Source: Pantheon v1.0 (Skiena et al.) enriched by Yu et al. with network connections. Edges reflect Wikipedia hyperlinks as of April 2018.

Node attributes: birth year/country/continent, gender, occupation domain (8 categories: Arts, Institutions, Science, Humanities, Sports, Business, Exploration, Public Figure), Historical Popularity Index (HPI), average Wikipedia views.

Key biases: Western-centric Wikipedia community; recency bias; modern country borders applied to historical figures.

Notable patterns: over half the dataset is European; the 20th century accounts for >50% of individuals; men outnumber women 6.6:1. HPI top-10: Aristotle, Plato, Jesus; in contrast to views top-10 (Kim Kardashian, Eminem, Justin Bieber).

Network Structure

Global properties: the network is large and sparse (density ≈ 0.001), with 98% of nodes in a single weakly connected component, average degree 22.2 (median 14), reciprocity 0.346, transitivity 0.156, and diameter 15.

Homophily: strong assortativity by domain (0.69) and birth century (0.61), moderate by continent (0.48), country (0.33), and gender (0.33). Sports figures are the most self-referential and isolated cluster; humanities figures attract disproportionate cross-domain links.

Centrality indices compared:

Index	Top node
Degree	Barack Obama
PageRank	Adolf Hitler
Betweenness	Pope John Paul II
Harmonic	Bob Dylan
Eigen-centrality	Roger Federer[†]
HPI	Aristotle

[†] distorted by ATP tennis cluster density

Degree and PageRank are strongly correlated; betweenness identifies structural bridges (e.g. Charlemagne, Pelé) not captured by degree alone. Eigen-centrality exhibits a tennis-cluster artifact, making it unreliable here.

Visualization: YifanHu and ForceAtlas2 layouts confirm domain-based community structure: layout emerges from topology alone, validating homophily findings.

Inference: ERGMs

ERGMs model the probability of an observed network as:

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

where $g(\mathbf{Y})$ are network statistics (edges, reciprocity, shared partners, homophily terms) and $\kappa(\theta)$ is an intractable normalizing constant requiring MCMC approximation.

Three subnetworks were fitted: **18th-century figures** ($n = 525$, dyadic independence), **Irish figures** ($n = 53$, Markov + gwesp), and a **pseudo-random subnetwork** ($n = 111$, Markov + geometrically weighted degree). Across all models, shared domain, birth century, continent, and gender significantly increase edge probability, consistent with the exploratory findings.

Key challenge: model degeneracy. Goodness-of-fit diagnostics reveal that all fitted models produce simulated networks with anomalous degree distributions (a few outlier networks inflate variance and pull means far from medians). Observed statistics consistently fall at the boundary of simulation intervals. The models capture density and reciprocity adequately but fail to reproduce the empirically observed democratic degree distribution.

Key Findings & Limitations

- **Homophily is the dominant structural force**, especially by occupation domain and historical period.
- **PageRank and degree agree** on globally important figures; betweenness surfaces structural bridges not identified by volume alone.
- **ERGMs are flexible but fragile**: degeneracy, sensitivity to statistic choice, and computational limits ($N < 1000$) constrain inference on large directed networks.
- **Gender findings are inconclusive** due to severe imbalance; women show higher out-degree but the dataset composition precludes strong conclusions.

Technologies

Languages & tools: R, `igraph`, `statnet`, `ergm`, `network`, `ggplot2`, Gephi.
Data: Pantheon v1.0 (Yu et al. network extension).

Future work: mixed membership stochastic block models, Pantheon 2.0 network extension, bootstrap/Bayesian ERGM estimation for larger subnetworks.