Assigment 1-Exploring Real-World Networks

Analysis of a Real-World Social Network: YouTube

November 9, 2025

tutor:
MH Lees

group:
MA Complex Systems and Policy

student:
Agaath de Vries
10205071

course:
Model Based Descision-making

course˙id:
5404MBDM6Y

A. Dataset Selection and Description

Dataset: YouTube Social Network
Source: Stanford Network Analysis Project (SNAP)
Context: This expansive dataset illustrates the friendships between users on YouTube. The graph is undirected, enabling the examination of its topology to reveal insights into online social structures, information dissemination, and community formation.
Loading Methodology: Data is extracted from the compressed edge list file `com-youtube.ungraph.txt.gz`, utilizing Python's `networkx` library. Since the graph consists solely of the Largest Connected Component (LCC), analyses are conducted on the complete network.

Table 1: Descriptive Statistics of the YouTube LCC

Metric	Interpretation	Value
Nodes (N)	Total users analyzed.	1,134,890
Edges (E)	Total friendships.	2,987,624
Average Degree ($\langle k \rangle$)	Average number of connections per user.	5.2651
Density (ρ)	Connectedness relative to a complete graph.	$\approx 4.64 \times 10^{-6}$
Average Clustering Coefficient (C)	Measure of local community formation.	0.0808
Avg. Path Length (L)	Global efficiency (proxy value).	≈ 6.5
Degree Assortativity (r)	Preference for connecting to nodes of similar degree.	0.0036

B. Global Network Properties and Pre-processing

The network under consideration is characterized by an extremely low density, with the connection probability ρ approximately equal to 4.64×10^{-6}. This translates to a very sparse network structure. An analysis of the Average Clustering Coefficient, denoted as C and valued at 0.0808, reveals that while it may seem modest when considered in isolation, it is considerably higher than what one would anticipate in a random graph with an equivalent density, where the expected clustering coefficient C_{ER} is approximately equal to ρ. This disparity highlights the presence of non-random, localized grouping patterns, which serves as evidence of discernible social clusters or circles within the network platform.

Further examination of the network's topology is provided by the Average Shortest Path Length, L, which is approximated by the 90th percentile effective diameter as supplied by the SNAP dataset, yielding a value of approximately 6.5 **snap˙youtube**. Such a small value underpins the **small-world** phenomenon, suggesting that the vast majority of users within the network are merely a few connections apart. The Degree Assortativity, represented by $r = 0.0036$, approaches zero, thus indicating a non-assortative network structure in which there is no preferential attachment between network hubs.

While the average shortest path length ($L \approx 6.5$) confirms the "small-world" phenomenon, the averageand the clustering coefficient ($C = 0.0808$) is exceptionally low). This combination is particularly insightful. The formal metric of the 'small-world' phenomenon, termed sigma (σ), provides a quantitative assessment by comparing the network's clustering coefficient (C) and path length (L) against those of a random (Erdős-Rényi graph, with C_{rand} and L_{rand} denoting their respective metrics Watts **and** Strogatz 1998. The sigma is computed as:

$$\sigma = \frac{C/C_{rand}}{L/L_{rand}}$$

In the context of this network, $L \approx 6.5$ is significantly shorter than the path length of an equivalent random graph, which is approximately $L_{rand} \approx 14.5$. Consequently, the ratio L/L_{rand} is considerably less than one. In contrast, the clustering coefficient $C = 0.0808$ is *orders of magnitude* higher than that of the ER model, which stands at $C_{rand} \approx 4.64 \times 10^{-6}$. The resultant σ value, being significantly greater than one, provides empirical validation of the network's 'small-world' nature, despite the social clustering not being as high as in networks such as Facebook.

C. Comparative Analysis: Real vs. Model Networks

The YouTube network was compared against three classical benchmark models parameterised to match the real network's size (N) and average degree ($\langle k \rangle \approx 5.2651$). The Average Path Length (L) values for all large networks are computed in the notebook using a **sampling approximation** based on 10,000 source nodes.

1. **Erdős-Rényi (ER) Model** ($\mathbf{G(N, p)}$): Fails to reproduce the high clustering and the structural variation in node degrees.
2. **Watts-Strogatz (WS) Model** ($\mathbf{WS(N, k, p)}$): Successfully predicts the small-world property (low L), but fails catastrophically on the degree distribution, predicting a peaked distribution instead of a heavy tail.
3. **Barabási-Albert (BA) Model** ($\mathbf{BA(N, m)}$): Successfully predicts the heavy-tailed degree distribution and the low absolute clustering coefficient, strongly matching the observed large-scale topology.

The confirmation that the YouTube network follows a Barabási **and** Albert 1999 (BA) model has direct implications for its resilience, a concept explained by **percolation theory**. This theory examines how network integrity is maintained as nodes are removed. Scale-free networks, unlike their ER and WS counterparts, exhibit a unique 'robust-yet-vulnerable' nature Albert, Jeong **and** Barabási 2000. They are highly **robust** to random failures; the removal of random, low-degree nodes (the vast majority of users) has a negligible effect on the network's connectivity. However, they are extremely **vulnerable** to targeted attacks. Because the network's connectivity is maintained by a small number of high-degree hubs, a targeted removal of these top-degree nodes would quickly shatter the network into disconnected fragments, catastrophically disrupting its function.

D. Centrality and Node Importance

Due to the graph size ($N > 1$M), exact computation of path-dependent metrics (Closeness and Betweenness) is computationally intractable. To provide the required measures, the notebook employs **large-scale approximation via sampling** ($k = 10,000$) for these centralities **newman˙networks**. The confirmation

Table 2: Comparative Network Metrics

Metric	Real Network (LCC)	Erdős-Rényi (ER)	Watts-Strogatz (WS)	Barabási-Albert (BA)
Avg. Clustering (C)	0.0808	$\approx 4.64 \times 10^{-6}$	0.5001	0.0076
Avg. Path Length (L)	≈ 6.5 (eff. diameter)	≈ 14.5 (analytic)	≈ 7.3	≈ 7.1
Assortativity (r)	0.0036	-0.0001	0.0007	-0.0051
Degree Distribution	Power-Law (Heavy Tail)	Poisson (Peaked)	Peaked	Power-Law

that the YouTube network follows a Barabási **and** Albert 1999 (BA) model has direct implications for its resilience, a concept explained by **percolation theory**. This theory examines how network integrity is maintained as nodes are removed. Scale-free networks, unlike their ER and WS counterparts, exhibit a unique 'robust-yet-vulnerable' nature Albert, Jeong **and** Barabási 2000. They are highly **robust** to random failures; the removal of random, low-degree nodes (the vast majority of users) has a negligible effect on the network's connectivity. However, they are extremely **vulnerable** to targeted attacks. Because the network's connectivity is maintained by a small number of high-degree hubs, a targeted removal of these top-degree nodes would quickly shatter the network into disconnected fragments, catastrophically disrupting its function. Due to the graph size ($N > 1$M), exact computation of path-dependent metrics (Closeness and Betweenness) is computationally intractable. To provide the required measures, the notebook employs **large-scale approximation via sampling** ($k = 10,000$) for these centralities **newman˙networks**.

1. **Degree Centrality:** Measures local popularity (number of friends). Top nodes are local hubs with maximum direct connections.
2. **Eigenvector Centrality:** Measures influence derived from connections to other highly connected nodes **newman˙networks**. Top nodes are the true 'influencers' embedded deep within the central parts of the social network.
3. **Betweenness Centrality (Sampled):** Approximates control over information flow **newman˙networks**. Top nodes represent critical **bridges** or gatekeepers connecting disparate communities.
4. **Closeness Centrality (Sampled):** Approximates efficiency in reaching all other nodes **newman˙networks**. Top nodes are the most centrally located and efficient communicators.

E. Discussion and Interpretation

Comparison Summary: The YouTube social network is a **Scale-Free Small-World Hybrid Network**. It is best modeled by the Barabási-Albert model **barabasi˙science** due to the observed power-law degree distribution and low absolute clustering.

Real-World Mechanisms: The structure is a direct result of two competing forces **barabasi˙science**:

- **Preferential Attachment:** As new users join, they are far more likely to "friend" highly visible, established users (hubs), driving the "rich-get-richer" dynamic that creates the heavy-tailed degree distribution.

- **Weak Triadic Closure:** The low clustering indicates that friends of a content creator (hub) are not likely to be friends with each other, reinforcing the idea that connections are based on shared interest in a central node, not mutual acquaintance.

The degree assortativity coefficient of $r \approx 0.0036$ is a particularly revealing metric. This value is effectively zero, indicating the network is **non-assortative**—there is no statistical preference for nodes to connect to others of a similar degree Newman 2003. This neutrality is significant because it contrasts sharply with the two common archetypes.Highly **assortative** networks ($r > 0$) are typical of reciprocal social networks (like Facebook), where 'popular' users tend to be friends with other popular users. Conversely, highly **disassortative** networks ($r < 0$) are common in technological or biological systems (like the Internet or protein-interaction networks), where high-degree hubs (e.g., routers) primarily connect to many low-degree nodes (e.g., end-users).The YouTube network's neutrality suggests a hybrid mechanism. While it functions as a social network, its 'friendship' model is not purely reciprocal. It is dominated by a 'subscriber' dynamic, where many low-degree users connect to high-degree content creators (a disassortative force). This is likely balanced by a simultaneous assortative force, where high-degree creators 'friend' or collaborate with other high-degree creators. The result is a network where these competing mixing patterns effectively cancel each other out, producing a neutral assortativity coefficient. **Implications:**

- **Robustness vs. Vulnerability:** The scale-free structure provides high robustness against random failures (low-degree nodes deleting accounts) but results in extreme vulnerability to targeted attacks or failures affecting the top Degree and Betweenness hubs **barabasi˙science**. The system relies entirely on a few critical nodes for global connectivity.
- **Rapid Dissemination:** The small-world property ($L \approx 6.5$) ensures that trends, content, and information can spread across the entire user base with exceptional speed.

References

Albert, Réka, Hawoong Jeong **and** Albert-László Barabási (2000). **?**Error and attack tolerance of complex networks**?** **in***Nature*: 406.6794, **pages** 378–382.
Barabási, Albert-László **and** Réka Albert (1999). **?**Emergence of scaling in random networks**?** **in***Science*: 286.5439, **pages** 509–512.
Newman, Mark EJ (2003). **?**Mixing patterns in networks**?** **in***Physical Review E*: 67.2, **page** 026126.
Watts, Duncan J. **and** Steven H. Strogatz (1998). **?**Collective dynamics of 'small-world' networks**?** **in***Nature*: 393.6684, **pages** 440–442.