Master in Data Science

Complex and Social Networks

The Role of Network Topology, Seeding Strategy and Initial Adopters in Political Information Diffusion On Social Media

Ilaria Boschetto

ilaria.boschetto@estudiantat.upc.edu

Gabriele Villa

gabriele.villa@estudiantat.upc.edu

Project Report

Academic Year 2025/2026

1 Introduction

Understanding how information propagates through social networks is a fundamental challenge in network science with profound implications for viral marketing, political campaigns, public health interventions, and the diffusion of misinformation. Social media platforms such as Facebook, Twitter, and Instagram have transformed traditional communication patterns, enabling content to reach millions of users within hours through cascading effects across network connections. In these digital ecosystems, the structural properties of networks, such as degree distribution, clustering, and centrality, interact with behavioral mechanisms to determine both the speed and extent of information diffusion.

This study investigates information diffusion in Facebook page-page interaction networks, where nodes represent Facebook pages and edges represent mutual likes or follow relationships. We focus our analysis on the politician network, which comprises Facebook pages of political figures and organizations: understanding information spread in political networks has direct implications for electoral campaigns, policy communication, and the spread of political misinformation.

Our research addresses three interrelated questions:

- **First question:** does the topological structure of real social networks significantly affect diffusion dynamics compared to random graph models with equivalent basic characteristics?
- **Second question:** how does the selection of initial adopters influence the speed and final reach of information cascades?
- **Third question:** do different diffusion models produce significantly different outcomes in terms of speed and reach, reflecting distinct mechanisms of social influence?

To address these questions, we compared diffusion dynamics on the real politician network against two random graph benchmarks:

- Erdős-Rényi graphs (which preserve network size and density)
- Degree-preserving switching models (which additionally maintain the degree distribution).

We implemented three standard diffusion models, each capturing different mechanisms of social influence:

- the Independent Cascade Model (ICM)
- the Linear Threshold Model (LTM)
- the Susceptible-Infected-Recovered (SIR) model

In addition, for each network-model combination, we evaluated three seeding strategies based on node centrality measures (degree and betweenness centrality) as well as random selection.

Finally, we tested three hypotheses:

- **Hypothesis 1:** we hypothesize that the speed of information diffusion differs significantly between real social media networks and random benchmarks, even when controlling for basic network statistics. This would indicate that realistic topological features such as clustering, community structure, and assortativity play a crucial role in shaping diffusion dynamics.
- **Hypothesis 2:** we hypothesize that centrality-based seeding strategies significantly outperform random seeding in terms of both diffusion speed and final cascade size. This would provide quantitative evidence for the importance of targeting influential nodes in viral campaigns and interventions.
- **Hypothesis 3:** we hypothesize that the choice of diffusion model has a significant effect on both the speed and final reach of information spread. This would reflect fundamental differences in how independent cascades, threshold-based adoption, and epidemic dynamics with recovery capture real-world information propagation mechanisms.

By combining network analysis with computational simulations and statistical testing, we systematically evaluate how network structure, seed selection, and model choice affect information diffusion. Our findings lay the groundwork for developing actionable insights that could inform political communication strategies and public awareness campaigns, while contributing to the broader understanding of information spread in social networks and providing a foundation for more comprehensive analyses in this domain.

2 Results

2.1 Network statistics

We first present some basic statistics for all the networks included in the SNAP database:

Network	Nodes	Edges	Density	AvgDegree	AvgClustering	Diameter	Assortativity	LCC Size
artist	50515	819090	0.0006	32.4296	0.1472	11	-0.0191	50515
politician	5908	41706	0.0024	14.1185	0.4286	14	0.0182	5908
athletes	13866	86811	0.0009	12.5214	0.3033	11	-0.0270	13866
company	14113	52126	0.0005	7.3869	0.2872	15	0.0126	14113
government	7057	89429	0.0036	25.3448	0.4326	10	0.0293	7057
new_sites	27917	205964	0.0005	14.7555	0.3199	15	0.0219	27917
public_figure	11565	67038	0.0100	11.5933	0.2149	15	0.2020	11565
tvshow	3892	17239	0.0023	8.8587	0.4433	20	0.5605	3892

Table 1: Facebook Network Statistics

We chose to analyze the politician network: in Figure 1 we show its degree distribution both in linear scale and in log-log scale for clarity.

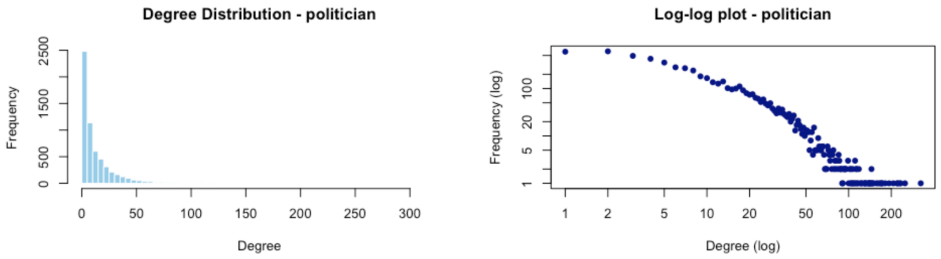

Figure 1: Degree distribution of the politician network (linear scale and log-log scale)

Figure 2 visualizes the politician network structure with hub nodes highlighted according to two centrality measures. The visualization is based on a sample of 500 nodes from the largest connected component. The left panel identifies hub nodes as those in the top 5% of degree centrality (highest number of direct connections), while the right panel highlights hubs based on betweenness centrality (nodes lying on many shortest paths). Hub nodes are colored in red and sized larger than regular nodes (grey) to emphasize their structural importance in the network.

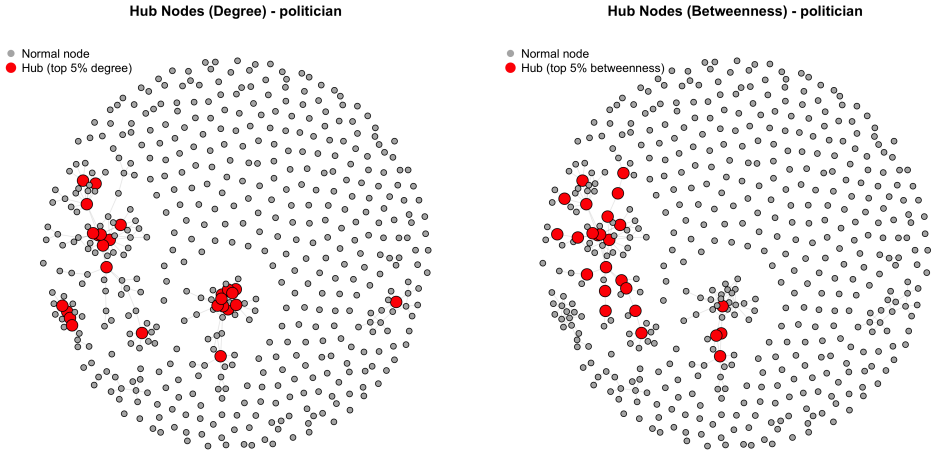

Figure 2: Degree distribution of the politician network (linear scale and log-log scale)

2.2 Results of the simulations

We begin by examining final cascade sizes across all experimental conditions. The following three figures (Figure 3, Figure 4 and Figure 5) present results for each diffusion model separately, with each figure showing three panels corresponding to the three network types. Within each panel, boxplots compare the three seeding strategies, revealing how network structure and seed selection jointly influence information reach for each propagation mechanism.

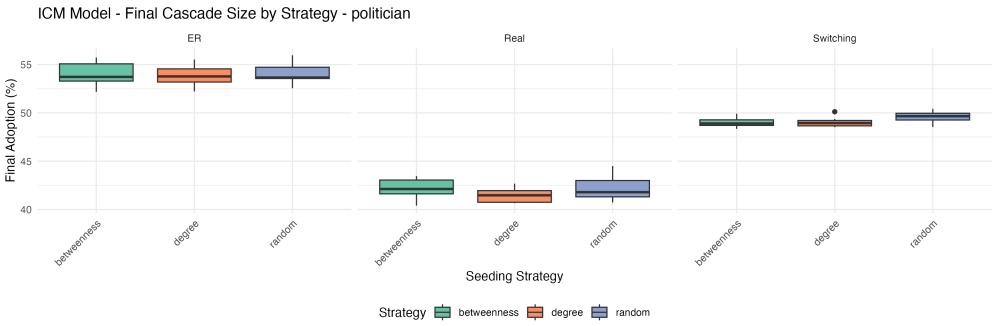

Figure 3: Final cascade size by strategy for the ICM diffusion model

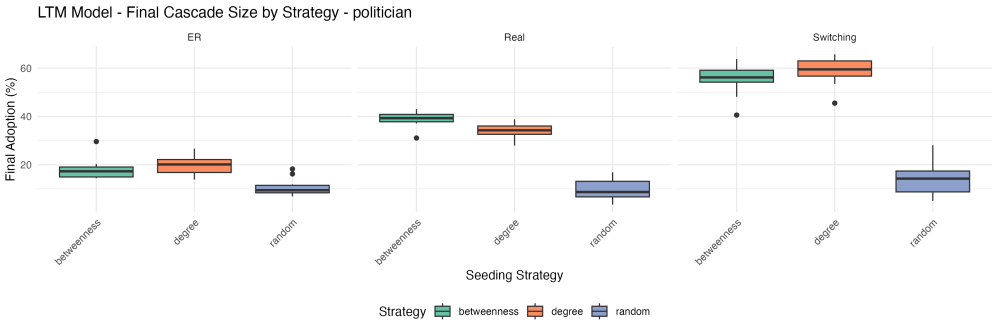

Figure 4: Final cascade size by strategy for the LTM diffusion model

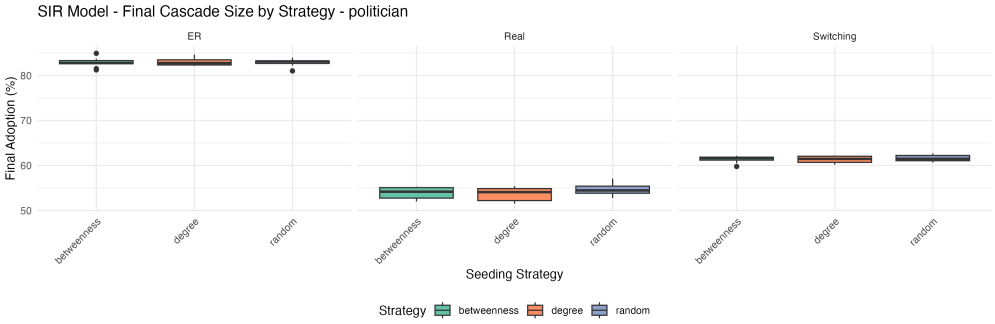

Figure 5: Final cascade size by strategy for the SIR diffusion model

We next examined diffusion speed across all experimental conditions. Figure 6, 7 and 8 show the number of time steps to convergence for each diffusion model across all combinations of network type and seeding strategy. As with the previous analysis, each figure contains three panels representing the real network, Erdős-Rényi benchmark, and switching model, with boxplots comparing the three seeding strategies. This complementary perspective reveals how quickly information cascades develop under different structural and strategic conditions.

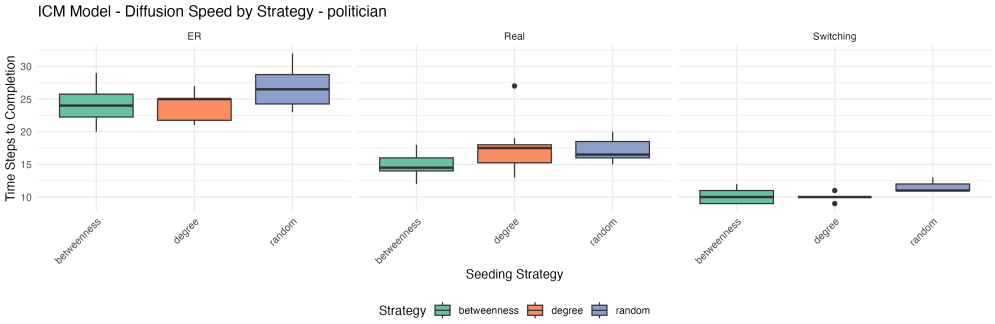

Figure 6: Diffusion speed by strategy for the ICM diffusion model

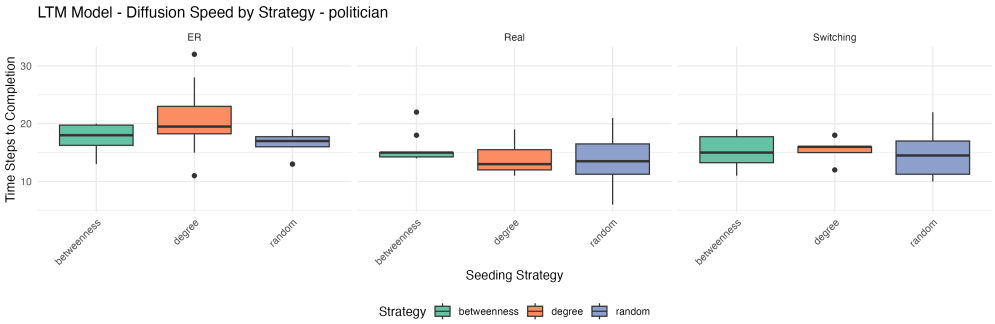

Figure 7: Diffusion speed by strategy for the LTM diffusion model

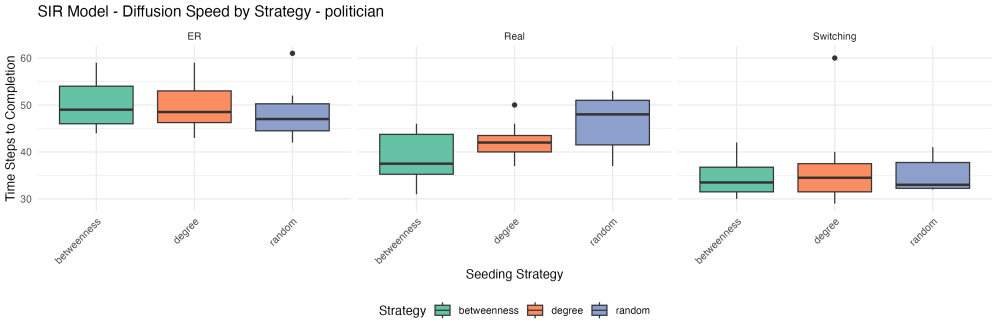

Figure 8: Diffusion speed by strategy for the SIR diffusion model

Figure 9 shows the cumulative number of adopted nodes over time for all three diffusion models on a single simulation run using the real politician network with degree centrality seeding. The y-axis displays adoption rates as percentages of the total network.

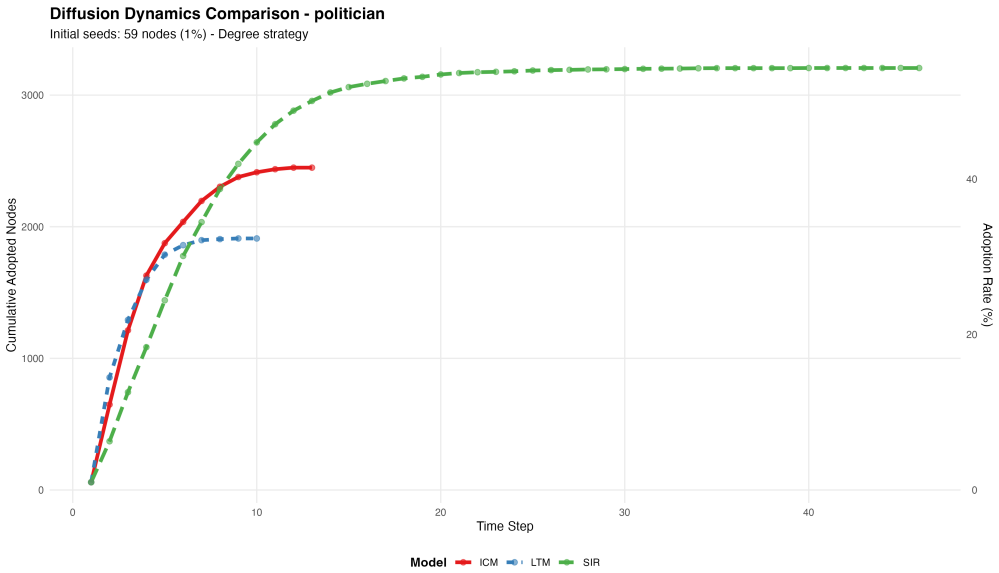

Figure 9: Cumulative adopted nodes over time for the three diffusion models

Figure 10 displays the SIR model's compartmental dynamics, showing the number of nodes in each state (Susceptible, Infected, Recovered) over time. The vertical dashed line marks the peak of infections.

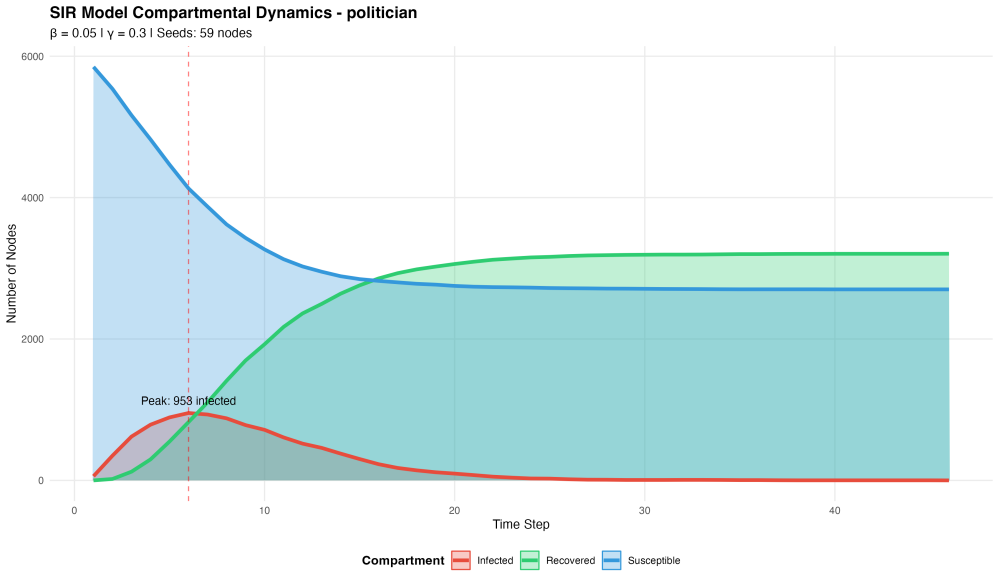

Figure 10: SIR model compartmental dynamics showing the evolution of Susceptible, Infected, and Recovered populations over time using degree-based seeding strategy

2.3 Hypothesis testing

Here we present the results of the three hypothesis tests.

2.3.1 Hypothesis 1

To test the first hypothesis, we compared diffusion speed between the real politician network and the two random benchmarks using independent samples t-tests. For each diffusion

model, we performed pairwise comparisons between the Real network and the Erdős-Rényi benchmark, and between the Real network and the Switching model. Table 2 presents the results of these statistical tests, including t-statistics, degrees of freedom, p-values, and 95% confidence intervals for the mean difference in convergence times. Significant p-values ($<$ 0.05) indicate that diffusion speed differs meaningfully between network types. Figure 11 displays the distribution of convergence times (number of steps until stopping) for each network type across all three diffusion models.

Model	Comparison	t	df	p-value	95% CI Lower	95% CI Upper
ICM	Real vs ER	-11.586	57.932	$< 2.2 \times 10^{-16}$	-10.086	-7.114
ICM	Real vs Switching	10.653	37.660	6.38×10^{-13}	4.779	7.021
LTM	Real vs ER	-3.831	54.311	0.0003	-5.687	-1.780
LTM	Real vs Switching	-1.016	56.995	0.314	-2.377	0.777
SIR	Real vs ER	-4.755	57.072	1.39×10^{-5}	-9.616	-3.917
SIR	Real vs Switching	4.767	58.000	1.30×10^{-5}	4.177	10.223

Table 2: Summary of Welch Two Sample t-tests across all models

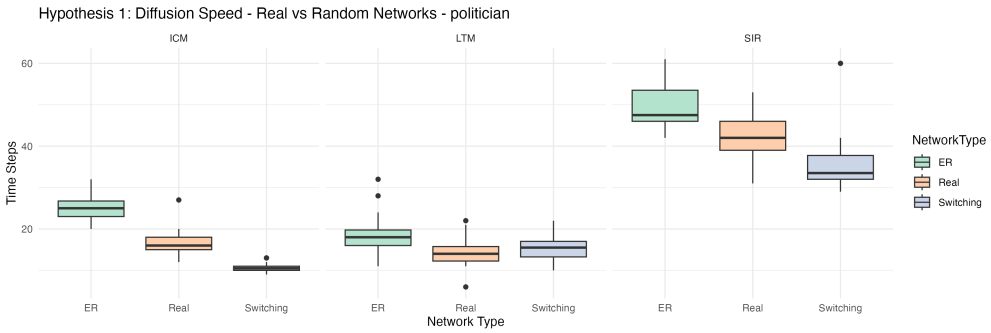

Figure 11: Steps needed for the algorithms to stop for each netowrk type for the three diffusion models

2.3.2 Hypothesis 2

To test the second hypothesis, we performed one-way ANOVA to assess whether seeding strategy significantly affects both diffusion speed and final cascade size on the real politician network. For each diffusion model, we tested whether the three seeding strategies produced different outcomes. When ANOVA revealed significant differences ($p < 0.05$), we conducted post-hoc Tukey HSD tests to identify which specific strategy pairs differed significantly while controlling for multiple comparisons. Figure 12 displays the distribution of convergence times for each seeding strategy across all three diffusion models.

Source	Df	Sum Sq	Mean Sq	F value	p-value
ICM	2	38.07	19.033	2.657	0.088
Residuals	27	193.40	7.163		

Table 3: ANOVA: Seeding Strategy Effect on ICM Speed

Source	Df	Sum Sq	Mean Sq	F value	p-value
LTM	2	22.87	11.43	1.092	0.35
Residuals	27	282.60	10.47		

Table 4: ANOVA: Seeding Strategy Effect on LTM Speed

Source	Df	Sum Sq	Mean Sq	F value	p-value
SIR	2	298.5	149.23	5.798	0.008**
Residuals	27	694.9	25.74		

Table 5: ANOVA: Seeding Strategy Effect on SIR Speed

Comparison	Diff	Lower CI	Upper CI	p-adj
Degree – Betweenness	3.3	−2.325	8.925	0.328
Random – Betweenness	7.7	2.075	13.325	0.006**
Random – Degree	4.4	−1.225	10.025	0.147

Table 6: Post-hoc Tukey HSD for SIR Speed

Source	Df	Sum Sq	Mean Sq	F value	p-value
ICM	2	11.475	5.737	1.632	0.214
Residuals	27	94.947	3.517		

Table 7: ANOVA: Seeding Strategy Effect on ICM Final Size

Source	Df	Sum Sq	Mean Sq	F value	p-value
LTM	2	17,235.835	8,617.918	189.3	$< 2 \times 10^{-16}$***
Residuals	27	1,229.330	45.531		

Table 8: ANOVA: Seeding Strategy Effect on LTM Final Size

Source	Df	Sum Sq	Mean Sq	F value	p-value
SIR	2	20.595	10.298	1.512	0.238
Residuals	27	183.825	6.808		

Table 9: ANOVA: Seeding Strategy Effect on SIR Final Size

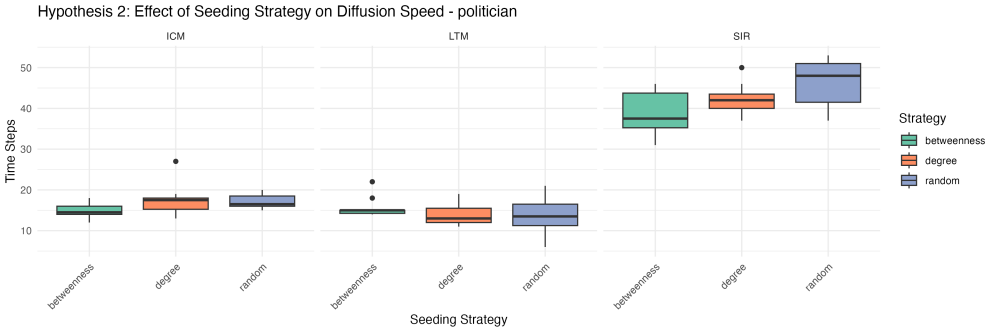

Figure 12: Steps needed for the algorithms to stop for each seeding strategy for the three diffusion models

2.3.3 Hypothesis 3

For the third hypthesis, we performed a one-way ANOVA to test whether different diffusion models produce significantly different cascade sizes when applied to the same network structure. This analysis helps us understand whether the choice of diffusion mechanism itself influences the spread outcomes, independent of network topology or seeding strategy. The ANOVA results are presented in Table 10, while pairwise comparisons between models are shown in Table 11.

Source	Df	Sum Sq	Mean Sq	F value	p-value
Model	2	6,846,140	3,423,070	241.6	$< 2 \times 10^{-16}$ ***
Residuals	27	382,608	14,171		

Table 10: ANOVA: Effect of Model on Final Reach

Comparison	Diff	Lower CI	Upper CI	p-adj
LTM – ICM	-440	-571.996	-308.004	< 0.001 ***
SIR – ICM	719	587.004	850.996	< 0.001 ***
SIR – LTM	1,159	1,027.004	1,290.996	< 0.001 ***

Table 11: Post-hoc Tukey HSD Comparisons for Model Effect

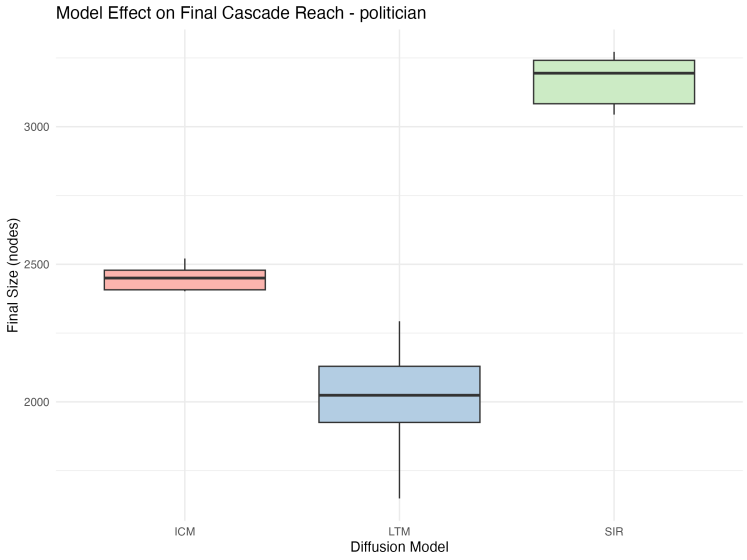

Figure 13: Number of nodes that adopted the information at the end of the spread for the different diffusion models for the degree-based seeding strategy

2.4 Final results

The following tables (Table 12, Table 13 and Table 14) summarize the best-performing seeding strategies identified for each diffusion model. For every network topology analyzed, we isolated the strategy that yielded the largest average cascade size.

NetworkType	Strategy	ICM_AvgSize	ICM_AvgSteps
ER	random	3193.1	26.8
Real	betweenness	2492.6	14.9
Switching	random	2929	11.6

Table 12: Best Performers for the ICM diffusion model

NetworkType	Strategy	LTM_AvgSize	LTM_AvgSteps
ER	degree	1159.9	20.6
Real	betweenness	2294.7	15.7
Switching	degree	3469.3	15.7

Table 13: Best Performers for the LTM diffusion model

NetworkType	Strategy	SIR_AvgSize	SIR_AvgSteps
ER	degree	4902.2	49.4
Real	random	3229.9	46.6
Switching	random	3641.3	35.1

Table 14: Best Performers for the SIR diffusion model

Table 15 presents the aggregated performance metrics for each diffusion model across the three network topologies. To derive these values, simulation results were grouped by network type to calculate the mean for two key indicators: *Reach*, defined as the average fraction of the total population reached by the diffusion and *Speed*, defined as the average number of time steps required for the diffusion process to complete

	ICM		LTM		SIR	
Network	**Reach(%)**	**Speed**	**Reach(%)**	**Speed**	**Reach(%)**	**Speed**
ER	53,93	25,067	16,17	18,200	82,91	493,33
Real	41,92	16,467	27,46	14,467	54,06	42,567
Switching	49,20	10,567	42,70	15,267	61,46	35,367

Table 15: Performance of every diffusion model for every network type

3 Discussion

3.1 Seeding Strategy Performance Across Network Topologies

The cascade size distributions across network topologies (Figures 3, 4, 5) reveal that seeding strategy effectiveness depends critically on both network structure and diffusion mechanism.

For the ICM and SIR models, seeding strategies yield nearly equivalent performance across all network types. In ICM, because each activation attempt is independent and probabilistic, the structural advantages of centrality-based seeds are largely masked by randomness, leading to performance differences typically below 2 percentage points. In the SIR model, the high recovery rate causes infected nodes to stop spreading information quickly, preventing any seeding strategy from achieving substantially larger cascades than others.

In contrast, the LTM model shows large differences between strategies, especially on the real network where betweenness centrality reaches around 40% adoption compared to around 35% for degree and only 12% for random seeding. This occurs because the threshold mechanism requires nodes to receive influence from multiple neighbors before adopting: bridge nodes can spread information across disconnected communities, while high-degree nodes may be trapped within local clusters.

3.2 Diffusion Speed Across Seeding Strategies

The diffusion speed results (Figures 6, 7, 8) reveal that seeding strategy has limited impact on convergence time compared to its substantial effect on final reach.

Across all three models, strategy differentiation in speed is minimal: differences typically range between 2-5 time steps regardless of network type. This contrasts sharply with the dramatic reach differences observed in the LTM model, where betweenness seeding achieved triple the adoption of random seeding on the real network. The LTM model shows particularly uniform convergence times (12-20 steps) across all conditions despite vastly different final cascade sizes, indicating that threshold-based cascades proceed at consistent rates: the critical factor is how far they spread, not how quickly they reach their stopping point.

The SIR model exhibits notably longer convergence times (30-50 steps) than ICM or LTM due to the gradual recovery process, and both in the SIR and in the ICM model the switching benchmark achieves faster diffusion.

3.3 Temporal Dynamics of Information Diffusion

Figure 9 illustrates the distinct temporal patterns of the three diffusion models through example cascades on the real politician network. The ICM and LTM models exhibit rapid initial growth followed by early saturation: ICM plateaus at around 2,500 nodes (about 42% adoption) after approximately 12 steps, while LTM converges even faster to around 1,900 nodes (about 32%) within 10 steps. In contrast, the SIR model shows sustained growth over a much longer period, reaching over 3,200 nodes (about 54%) and continuing to spread beyond 40 time steps.

3.4 SIR decomposition

Figure 10 decomposes the SIR process into its three compartments, revealing the characteristic epidemic curve. The infected population (red) peaks at 953 nodes around step 5, then gradually declines as recovery outpaces new infections. The recovered population (green) grows steadily throughout the simulation, ultimately representing the cascade's final reach. The susceptible population (blue) decreases rapidly in the first 10 steps but a substantial fraction (about 2,700 nodes) remains unreached by the end of the process. This pattern demonstrates how the balance between transmission ($\beta = 0.05$) and recovery ($\gamma = 0.3$) prevents complete saturation: rapid recovery causes local exhaustion of the cascade before it can reach all network regions, particularly isolated or peripheral nodes. This behavior mirrors real-world social media dynamics, where information rarely reaches the entire network and attention is inherently transient: news stories and viral content spread rapidly but lose relevance quickly as users move on to new topics, effectively "recovering" from exposure to previous content.

3.5 Analysis of Hypothesis 1

Our analysis reveals complex patterns in how network topology influences diffusion speed, with results varying substantially across different diffusion models. As shown in Figure 11 and Table 2, the real Facebook network exhibits model-dependent behavior compared to random benchmarks.

For ICM and SIR, the degree-preserving switching model shows the fastest diffusion (lowest time steps), followed by the real network, with the Erdős-Rényi random graph being the slowest. The real network is significantly faster than ER (ICM: $p < 2.2 \times 10^{-16}$, SIR: $p = 1.39 \times 10^{-5}$) but significantly slower than the switching model (ICM: $p = 6.38 \times 10^{-13}$, SIR: $p = 1.30 \times 10^{-5}$). This suggests that while degree distribution plays a crucial role in diffusion speed, the community structure and clustering present in the real network actually slow down global spread compared to a randomly rewired topology.

The LTM model shows a different pattern: the real network is faster than ER ($p = 0.0003$) but shows no significant difference compared to the switching model ($p = 0.314$). This indicates that for threshold-based diffusion, degree distribution alone explains diffusion speed, while the specific arrangement of connections has minimal impact.

3.6 Analysis of Hypothesis 2

We have seen that the impact of seeding strategy on diffusion speed varies considerably across models. For ICM in Table 3, we observe a marginally non-significant effect ($F = 2.657$, $p = 0.088$), while LTM (Table 4) shows no significant differences between strategies ($F = 1.092$, $p = 0.35$). However, the SIR model demonstrates a significant effect. The post-hoc Tukey HSD test in Table 6 reveals that Random seeding produces significantly slower diffusion than Betweenness seeding, with cascades completing approximately 7.7 time steps earlier ($p = 0.006^{**}$).

Regarding final cascade size, we observe a contrasting pattern. The LTM model exhibits an extremely strong effect of seeding strategy indicating that threshold-based diffusion is highly sensitive to the structural position of initial adopters. In contrast, neither ICM ($F = 1.632$, $p = 0.214$) nor SIR ($F = 1.512$, $p = 0.238$) show significant effects on final size.

These findings reveal that the SIR model is sensitive to seeding strategy regarding speed but not final coverage, suggesting different seeds may accelerate diffusion but ultimately reach similar population sizes. Conversely, the LTM model shows extreme sensitivity for final size but not speed, indicating certain seed positions enable information to cross critical thresholds and reach substantially larger populations. The ICM model appears most robust overall to seeding choices. We want to specify that these results remain contingent on our specific parameter choices.

3.7 Analysis of Hypothesis 3

The choice of diffusion model has a significant impact on the final cascade size, as demonstrated by the ANOVA results in Table 10 ($F = 241.6$, $p < 2 \times 10^{-16}$). As shown in Figure 13, the SIR model produces substantially larger cascades (median around 3,200 nodes) compared to both ICM (median around 2,450 nodes) and LTM (median around 2,000 nodes).

The Tukey HSD post-hoc comparisons in Table 11 reveal that all pairwise differences between models are statistically significant. The SIR model achieves the largest final reach, with cascades approximately 719 nodes larger than ICM ($p < 0.001$) and 1,159 nodes larger than LTM ($p < 0.001$). ICM produces intermediate cascade sizes, outperforming LTM by an average of 440 nodes ($p < 0.001$).

However, these comparisons between diffusion models should be interpreted with caution, as the relative performance depends heavily on the choice of parameters.

3.8 Conclusions and Future Work

Our analysis reveals that the three factors examined (network topology, seeding strategy, and diffusion model) have markedly different levels of influence on information propagation. Network topology proved to be a significant factor with model-dependent effects. In the ICM and SIR models, the real Facebook network exhibited intermediate diffusion speeds, falling between the slower Erdős-Rényi benchmark and the faster degree-preserving switching model. This pattern suggests that while degree distribution accelerates spread, the community structure and clustering present in real networks actually slow it down. The LTM model showed a different behavior, where diffusion speed was determined primarily by degree distribution alone, with the specific arrangement of connections having minimal impact. Seeding strategy, in contrast, demonstrated surprisingly limited influence across most conditions. Only the SIR model showed significant speed differences between strategies. The ICM model appeared largely robust to seeding decisions for both speed and reach. The diffusion model itself emerged as the most decisive factor. The SIR model consistently generated the largest cascades, followed by ICM and then LTM, with all differences being statistically significant. These findings suggest that while network structure and model choice fundamentally shape diffusion outcomes, the selection of initial seeds may be less critical than commonly assumed in many practical contexts. Different parameter values could produce substantially different results and alter the ranking between models. In our analysis, we selected parameters that we considered most realistic for modeling information diffusion in social networks, but the comparison itself remains contingent on these specific choices.

Future work could focus on a more systematic exploration of parameter space to better understand how different parameter choices affect model performance and comparative rankings. Additionally, extending this analysis to other networks from the SNAP dataset would help determine whether our findings generalize across different social network structures or are specific to the political Facebook network examined here. Investigating additional net-

work types, such as other online social platforms, could provide broader insights into the mechanisms driving information diffusion in complex social systems.

4 Methods

4.1 Dataset and Network Selection

We analyzed publicly available Facebook page-page networks from the SNAP (Stanford Network Analysis Project) dataset collection [1]. Among the available networks (artists, athletes, companies, government organizations, news sites, politicians, public figures, and TV shows), we selected the **politician network** for our primary analysis. This choice was motivated by several factors: first, political content is inherently designed for rapid dissemination and engagement, making it an ideal substrate for studying cascade dynamics; second, the network exhibits realistic social media characteristics including high clustering and community structure along ideological lines; third, understanding information spread in political networks has direct relevance for electoral campaigns, policy communication, and the mitigation of political misinformation.

The politician network comprises 5,908 nodes (Facebook pages) and 41,729 undirected edges representing mutual likes or follow relationships. We preprocessed the network by removing multiple edges and self-loops using the `igraph` package in R, ensuring a simple undirected graph structure suitable for diffusion analysis.

4.2 Random Network Benchmarks

To isolate the effect of realistic network topology on diffusion dynamics, we constructed two types of random benchmark networks that preserve different statistical properties of the original graph while eliminating higher-order structural features.

4.2.1 Erdős-Rényi Random Graphs

For each real network $G = (V, E)$ with $n = |V|$ nodes and $m = |E|$ edges, we generated an Erdős-Rényi random graph with the same number of nodes and expected number of edges. The edge probability p was set to $p = m/\binom{n}{2}$, preserving network size and density while randomizing all other structural properties. This baseline allows us to test whether the mere density of connections, independent of their arrangement, determines diffusion outcomes.

4.2.2 Degree-Preserving Switching Model

To assess the role of degree heterogeneity independently from clustering and community structure, we employed a degree-preserving rewiring algorithm. Starting from the real network, we performed $10 \times m$ edge switches, where each switch randomly selects two edges

[1] `https://snap.stanford.edu/data/gemsec-Facebook.html`

(u, v) and (x, y) and, if feasible, replaces them with (u, y) and (x, v) while maintaining simple graph properties. This process preserves the exact degree sequence of the original network while randomizing higher-order correlations such as clustering coefficients and community structure. The switching model thus serves as an intermediate benchmark between the real network and the Erdős-Rényi model.

4.3 Diffusion Models

We implemented three standard diffusion models that capture distinct mechanisms of social influence and information propagation. All models operate in discrete time steps, which we interpreted as one-hour intervals. We set a maximum simulation horizon of 168 time steps (1 week) as a computational safeguard, though in practice most diffusion processes reached steady state well before this limit through model-specific convergence criteria.

4.3.1 Independent Cascade Model (ICM)

The Independent Cascade Model assumes that influence propagates through independent probabilistic "trials" along network edges. At each time step t, every newly activated node u attempts once to activate each of its inactive neighbors v with probability p. We set a uniform activation probability $p = 0.1$ across all edges, reflecting a 10% chance that exposure to activated content leads to adoption. This value balances computational tractability with realistic engagement rates observed on social media platforms, where typical post engagement rates range from 1% to 15% depending on content type and audience characteristics.

Once a node becomes active, it remains active throughout the simulation.

Stopping criterion: The simulation terminates when no newly activated nodes successfully activate any neighbors in a given time step (i.e., the set of newly active nodes is empty), indicating that the cascade has exhausted its potential to spread further, or when the maximum time horizon of 168 steps is reached.

4.3.2 Linear Threshold Model (LTM)

The Linear Threshold Model incorporates social reinforcement by requiring nodes to receive sufficient cumulative influence before adopting. Each node v is assigned a random threshold $\theta_v \sim \text{Uniform}(0, 1)$ representing its resistance to adoption. At each time step, an inactive node v becomes active if the fraction of its active neighbors exceeds its threshold:

$$\frac{\sum_{u \in N(v) \cap A(t)} w_{uv}}{\sum_{u \in N(v)} w_{uv}} \geq \theta_v$$

We set uniform edge weights $w_{uv} = 1/|N(v)|$, normalizing by the degree of the recipient node. This weighting scheme reflects the intuition that individuals with fewer social connections are more strongly influenced by each active neighbor. The random threshold

assignment ensures heterogeneity in adoption propensity across the population, capturing realistic variation in susceptibility to social influence.

Stopping criterion: The simulation terminates when no nodes change state in a given time step, indicating that all remaining inactive nodes have insufficient active neighbors to reach their thresholds and will never activate, or when the maximum time horizon is reached.

4.3.3 Susceptible-Infected-Recovered (SIR) Model

The SIR model extends basic contagion dynamics by incorporating a recovery mechanism, allowing nodes to transition from an active (infected) state to a recovered state where they no longer propagate information (representing a person who has adopted the information and cannot be re-infected). This captures the realistic phenomenon of attention decay, where users who initially share content eventually lose interest or become saturated.

At each time step, every infected node u attempts to infect each susceptible neighbor v with probability $\beta = 0.05$ and independently recovers with probability $\gamma = 0.3$. We chose these parameters to achieve a reproduction number $R_0 = \beta/\gamma \approx 0.17 < 1$, resulting in subcritical diffusion that does not reach full network saturation. The parameter $\beta = 0.05$ reflects a 5% probability of transmission per exposed neighbor per time step, while $\gamma = 0.3$ implies an average infectious period of approximately 3-4 time steps (hours).

Stopping criterion: The simulation terminates when no infected nodes remain in the network (all nodes have either recovered or remain susceptible), as no further state transitions are possible without infected nodes, or when the maximum time horizon is reached.

4.4 Seeding Strategies

To evaluate the impact of initial adopter selection on diffusion outcomes, we compared three seeding strategies that represent different levels of structural information and targeting sophistication. For each network, we selected seed sets comprising 1% of the total nodes (approximately 59 nodes in the politician network). In the context of our politician network, seeding about 59 political figures represents a plausible scale for early adopters or influencers who might first share political information before it spreads to the broader network.

4.4.1 Random Seeding

As a baseline, we randomly sampled seed nodes uniformly from the network without considering any structural properties. This strategy represents untargeted interventions or organic content initiation by arbitrary users. To account for stochastic variation, random seeds were resampled independently for each Monte Carlo replication.

4.4.2 Degree Centrality Seeding

Degree centrality seeding selects the k nodes with the highest degree (number of direct connections). This strategy targets "hubs" or highly connected individuals who have large immediate reach. Degree-based targeting is computationally inexpensive and widely used in practice, as it requires only local connectivity information. The degree centrality of node v is simply $d(v) = |N(v)|$.

4.4.3 Betweenness Centrality Seeding

Betweenness centrality identifies nodes that lie on many shortest paths between other node pairs, effectively serving as bridges or brokers in the network. Seeding high-betweenness nodes may facilitate information flow between otherwise disconnected communities. The betweenness centrality of node v is defined as:

$$B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest paths from s to t, and $\sigma_{st}(v)$ is the number of those paths passing through v. While more computationally expensive than degree centrality, betweenness captures global structural importance rather than merely local connectivity.

4.5 Simulation Procedure

We conducted Monte Carlo simulations to estimate expected diffusion outcomes and quantify uncertainty due to stochastic variation in model dynamics and (for random seeding) seed selection. For each combination of network type (Real, Erdős-Rényi, Switching), seeding strategy (Random, Degree, Betweenness), and diffusion model (ICM, LTM, SIR), we performed 10 independent replications with different random number seeds.

To manage computational complexity, we implemented parallel processing using the `foreach` and `doParallel` packages in R, distributing simulation tasks across multiple CPU cores. Each worker process received a complete specification of the network structure, model parameters, and seed set, then executed the diffusion simulation independently. We recorded two primary outcomes for each simulation: the final cascade size (total number of nodes ever activated or, in the case of SIR, the number of recovered nodes representing those who were infected at some point) and the number of time steps until convergence (when model-specific stopping criteria were met).

The total experimental design comprised 3 network types $\times$ 3 seeding strategies $\times$ 3 diffusion models $\times$ 10 replications = 270 individual simulations.

4.6 Statistical Analysis

We tested three primary hypotheses using inferential statistics on the simulation results.

4.6.1 Hypothesis 1: Network Structure Effect

Hypothesis: The speed of information diffusion differs significantly between real social media networks and random benchmark networks with equivalent basic statistics.

To test this hypothesis, we compared the number of time steps until convergence across the three network types (Real, Erdős-Rényi, Switching) for each diffusion model separately. We employed independent samples t-tests to assess pairwise differences between real and Erdős-Rényi networks, and between real and switching model networks. Tests were conducted separately for ICM, LTM, and SIR models to account for potential model-specific effects. Statistical significance was assessed at the $\alpha = 0.05$ level. This hypothesis evaluates whether realistic topological features such as clustering, community structure, and degree assortativity meaningfully impact diffusion speed beyond what would be expected from density and degree distribution alone.

4.6.2 Hypothesis 2: Seeding Strategy Effect

Hypothesis: The speed and extent of information spread depend significantly on the type of initial adopters, with centrality-based strategies outperforming random seeding.

We tested this hypothesis using one-way analysis of variance (ANOVA) to compare diffusion speed (time steps to convergence) and final cascade size across the three seeding strategies within the real network. Separate ANOVAs were conducted for each diffusion model. When the F-test indicated significant differences ($p < 0.05$), we performed post-hoc pairwise comparisons using Tukey's Honestly Significant Difference (HSD) test to identify which specific strategies differed from one another while controlling the family-wise error rate. This analysis quantifies the practical benefit of investing in targeted seeding based on network structure, as opposed to uniform random dissemination.

4.6.3 Hypothesis 3: Diffusion Model Effect

Hypothesis: The choice of diffusion model has a significant effect on both the speed and final reach of information spread, reflecting different underlying mechanisms of social influence.

To test this hypothesis, we compared diffusion outcomes across the three models (ICM, LTM, SIR) on the real network using degree centrality seeding to control for seed selection effects. We performed one-way ANOVA on both final cascade size and convergence time, treating the diffusion model as the independent variable. Post-hoc Tukey HSD tests identified pairwise differences between models. This analysis reveals whether mechanistic assumptions about influence propagation (independent trials vs. threshold-based adoption vs. epidemic dynamics with recovery) lead to substantially different predictions, which has implications for model selection in applied settings.

4.7 Selection of Best-Performing Strategies

To identify the most effective seeding strategy for each network type, we compared the performance of all three seeding strategies across all three diffusion models. For each combination of network type and diffusion model, we calculated the average final cascade size across the 10 Monte Carlo simulations. We then ranked the seeding strategies based on their average performance and selected the strategy that achieved the highest average cascade size as the best-performing strategy for that specific network-model combination. This approach allowed us to determine which seeding strategy maximizes information diffusion under different structural conditions and diffusion mechanisms.

4.8 Software and Reproducibility

All analyses were conducted in R using the following packages: `igraph` for network construction and analysis, `dplyr` and `tidyr` for data manipulation, `ggplot2` for visualization, and `foreach`/`doParallel` for parallel computation. We set a global random seed (`set.seed(42)`) before network generation and used per-simulation seeds for replicability of stochastic processes. Complete code and processed results are available to ensure reproducibility of our findings.