# MASTERARBEIT | MASTER'S THESIS

Titel | Title

Designing Link Recommendation Methods to Reduce Polarization


verfasst von | submitted by

Jonas Luca Quenzer BSc


angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

| | |
|---|---|
| Studienkennzahl lt. Studienblatt \| Degree programme code as it appears on the student record sheet: | UA 066 645 |
| Studienrichtung lt. Studienblatt \| Degree programme as it appears on the student record sheet: | Masterstudium Data Science |
| Betreut von \| Supervisor: | Ass.-Prof. Dott.ssa Dott.ssa.mag. Yllka Velaj PhD |

# Abstract

**Minimizing Polarization in Social Networks through Strategic Edge Addition**

This thesis addresses the problem of polarization in social networks, where users primarily interact with like-minded individuals, creating echo chambers that limit exposure to diverse viewpoints. We develop algorithms to reduce polarization by strategically adding new edges between different ideological communities. Our approach leverages the Independent Cascade Model (ICM) to model information diffusion. It formulates the polarization minimization problem as finding the optimal set of k edges to add to the network that maximizes cross-community information flow.

Our objective function is monotone and submodular, which allows us to develop efficient approximation algorithms. The experimental study evaluates different seed selection methods and edge addition techniques across real-world networks (Twitter Interaction Network for the US Congress, AVES-WEAVER-SOCIAL) and synthetic networks with varying homophily values. Our results demonstrate that our Custom modification strategy consistently outperforms baseline approaches in activation metrics across various network configurations and seed functions.

The findings indicate that strategic edge addition is particularly effective in networks with high homophily ($h = 0.8$), where community divisions are strongest. The Custom strategy maintains its advantage across networks of different sizes, with smaller networks showing steeper initial activation but earlier plateaus, while larger networks show more gradual but sustained improvement with increasing k values.

This work contributes to understanding polarization dynamics in social networks. It offers practical approaches for promoting diverse information exposure, potentially creating healthier online environments that support democratic discourse through increased viewpoint diversity and reduced confirmation bias.

# Kurzfassung


**Minimierung der Polarisierung in sozialen Netzwerken durch strategische Kantenhinzufügung**

In dieser Arbeit geht es um das Problem der Polarisierung in sozialen Netzwerken. Oft interagieren Nutzer fast ausschließlich mit Gleichgesinnten, wodurch Filterblasen entstehen, die den Austausch mit anderen Meinungen einschränken. Um dem entgegenzuwirken, entwickeln wir Algorithmen, die gezielt neue Verbindungen zwischen unterschiedlichen ideologischen Gruppen schaffen. Dadurch soll die Polarisierung reduziert und der Austausch über verschiedene Gemeinschaften hinweg verbessert werden.

Unser Ansatz basiert auf dem Independent Cascade Model (ICM), das beschreibt, wie Informationen in einem Netzwerk verbreitet werden. Wir formulieren das Problem so, dass wir die optimalen k neuen Kanten identifizieren, die dem Netzwerk hinzugefügt werden sollten, um den Informationsfluss zwischen Gruppen zu maximieren.

Da unsere Zielfunktion monoton und submodular ist, können wir effiziente Approximationsalgorithmen entwickeln. In unseren Experimenten testen wir verschiedene Methoden zur Auswahl von Startknoten (Seeds) und zur Hinzufügung neuer Kanten. Diese Tests führen wir sowohl an realen Netzwerken – wie dem Twitter-Interaktionsnetzwerk des US-Kongresses und AVES-WEAVER-SOCIAL – als auch an synthetischen Netzwerken mit unterschiedlichen Homophilie-Werten durch. Unsere Ergebnisse zeigen, dass unsere speziell entwickelte Strategie durchweg bessere Ergebnisse liefert als herkömmliche Ansätze. Sie erreicht höhere Aktivierungswerte über verschiedene Netzwerkstrukturen und Seed-Strategien hinweg.

Besonders effektiv ist das strategische Hinzufügen von Kanten in Netzwerken mit starker Homophilie (h = 0,8), wo die Gruppenstrukturen am deutlichsten ausgeprägt sind. Unsere Methode zeigt ihre Stärken unabhängig von der Netzwerkgröße: Während kleinere Netzwerke anfangs einen schnellen Anstieg der Aktivierung verzeichnen, dieser aber früh abflacht, entwickeln größere Netzwerke eine gleichmäßigere, aber nachhaltige Verbesserung mit steigender Anzahl neuer Kanten.

Diese Arbeit hilft dabei, die Dynamik der Polarisierung in sozialen Netzwerken besser zu verstehen und bietet praktische Lösungen, um die Vielfalt der Meinungen in Online-Diskussionen zu fördern. Damit könnte sie einen Beitrag zu gesünderen digitalen Räumen leisten, die den offenen Diskurs durch mehr Meinungsvielfalt unterstützen.

# Contents

# 1 Introduction

In the past decade, social media and social networks have become increasingly popular amongst all ages in our society. Social media has become a prominent part of how individuals interact and are exposed to content and information. This content can be simple pictures of your friends, or it can be about a topic that might change or influence your opinion. The concern about influential platforms extends beyond social media and also includes search engines, which, as highlighted by David Shultz in a 2016 Science article [31], possess the potential to slightly introduce a bias to their user preferences and shape electoral outcomes by prioritizing certain content over others in search results. This underscores the unseen influence these platforms can exert on democratic processes, raising crucial questions about the transparency and fairness of algorithms controlling such vast information flows.

At the heart of this influence are recommendation systems, which are used by these platforms to decide what content is shown to users. Recommendation systems are powerful tools that aim to tailor content to individual preferences, making the user experience more engaging and relevant. Businesses rely on these systems to increase and maximize their revenue, optimizing different mathematical objectives to achieve this—whether by improving the retention rate of users, increasing the click-through rate, or recommending the product a user is most likely to buy.

However, considering how influential and powerful the big tech companies have become over the past few years, one should slightly shift one's viewpoint and think critically about the drawbacks these recommendation systems induce. Having the ability to influence the perspective of users should be viewed very critically, and recommendation systems should also aim at creating a more diverse and open online environment.

This paper will focus on the phenomenon called polarization. Polarization refers to the development of extreme and opposing viewpoints within a society or community. In social networks, polarization can create so-called echo chambers or filter bubbles. An echo chamber or filter bubble is created when individuals are mainly connected to like-minded citizens or information sources that support similar or the same opinions as their own. As a result, individuals within an echo chamber or filter bubble are not exposed to diverse content and information or information they might disagree with. Effectively they are isolated within their cultural and ideological bubble.

For example, people who supported the vaccination during COVID-19 followed and interacted mainly with information sources that shared their views. They might also create an online forum or community with like-minded people to exchange even more information within it. On the other hand, people who were against the vaccination founded their own groups and followed like-minded people. So these two groups coexisted with little to no exchange between them.

When users continuously engage with content that mirrors their own opinions, they increase polarization. The polarization increases because such environments reduce the chances of encountering diverse viewpoints, thus limiting the user's openness to opposing ideas. Furthermore, as polarization deepens, it accelerates the spread of misinformation and the adoption of extreme viewpoints as well as creating a division that hinders civil discourse and worsens social splits.

All in all, polarization reduces the information diversity and information exposure within a social network which increases the confirmation bias and hinders democratic discourse. So it is important to minimize polarization within a social network to create a healthier online environment that is more inclusive and supports diverse information diffusion.

The focus of this paper lies in designing an algorithm that fosters diversity in network connections while minimizing polarization. The algorithm aims to strategically add links to a network, creating new connections between communities with differing ideologies. The objective is to identify the best $k$ edges to add to the existing network to reduce polarization and increase the likelihood that users are exposed to a wider range of opinions.

To model a social network, we consider a directed, weighted graph in which the nodes represent users and the edges reflect their relationships. To illustrate opposing opinions, nodes are assigned one of two colors, representing the differing viewpoints within the network, as illustrated in Figure 1.1. The visualization demonstrates how a polarized network consists of distinct communities with limited connections, forming what we commonly refer to as "filter bubbles" or "echo chambers."

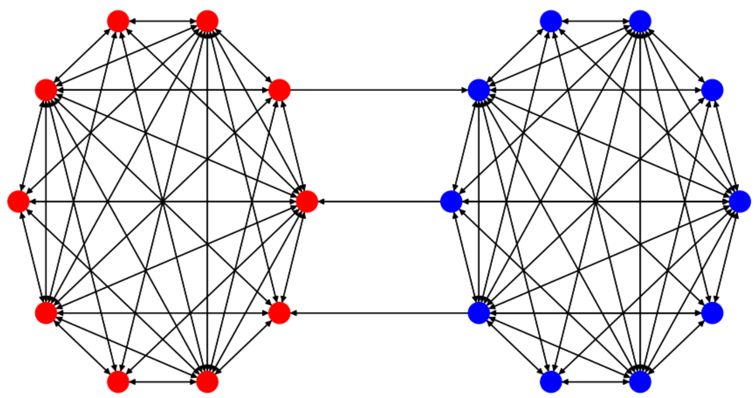

Figure 1.1: Example of Network with two filter bubbles, where the two different colors represent different opinions

For modeling the information dynamics within a social network we will use some diffusion models, namely the Independent Cascade Model (ICM, [3]). This diffusion model is designed to model the spread of some information throughout a system and is a good choice for modeling graph behavior in the context of this paper. While other works have addressed polarization and information diffusion in networks [28, 32, 24], they have often overlooked these specific diffusion models. Many have either used alternative models that do not capture the scenario as realistically as ICM, or avoided diffusion models altogether [23, 15]. By grounding our study in this diffusion model, we aim to provide a more realistic representation of real-world behavior.

Existing research has extensively explored the phenomenon of polarization on social media and proposed various methods to reduce it. Notably, several works have aimed at modifying the structure of a graph by adding new edges and using representations of nodes for users and edges for connections. For instance, Zhu et al. [32] developed algorithms based on the Friedkin-Johnsen Model to minimize polarization and disagreement within social networks through link recommendation systems. However, this approach uses a non-submodular monotonically increasing function, whereas our approach leverages the Independent Cascade Model (ICM) and Linear Threshold Model (LTM) and a different, submodular, and monotone objective function.

Similarly, Musco et al. [28] focused on structuring social networks to balance polarization and disagreement by recommending links between like-minded individuals or those with differing opinions. In contrast, their approach aims to suggest a set of connections to form a graph, rather than creating new connections within an existing graph, as we are proposing.

Other researchers, such as Haddadan et al. [24], have addressed polarization in networks resembling the Internet, where nodes represent web pages and edges represent links between them. Their work introduced metrics like the polarized bubble radius to quantify structural bias, but their focus was on individual user interactions rather than the broader network dynamics we consider.

Moreover, Corò et al. [12] examined link addition for maximizing influence in social networks, aiming to create well-connected and balanced networks. While they provided proof for submodularity and monotonicity in their objective functions using the ICM and LTM models, our research adapts these principles specifically to minimize polarization within social networks.

### Thesis structure

Chapter 2 presents a comprehensive review of related work in the field of polarization minimization and network analysis. Chapter 3 introduces the theoretical foundations and preliminary concepts, including formal definitions of the problem and a discussion of its computational complexity. Chapter 4 covers our experimental study methodology, including the diffusion process, seed selection methods, and edge addition techniques. Chapter 5 provides an extensive evaluation of our proposed methods across different network configurations, comparing performance metrics and analyzing the results. Finally, Chapter 6 concludes the thesis with a summary of our findings, limitations of our approach, and directions for future research.

### Used tools

We want to state that the author utilized Claude for text reformatting and generating Python code snippets to create graphics.

# 2 Related work

With the rise of social media also the research in this field also grew. Especially in recent years, different works have studied the phenomena of polarization on social media and how to reduce it. In this chapter, we will introduce some of their work, their key contributions, and the differences from our work.

In this chapter, we focus on research addressing polarization in social networks by modifying graph structures. A common approach involves strategically adding edges to bridge polarized communities, enhancing connectivity, and reducing polarization.

Papers with the highest similarity to our paper will be the ones that aim at adapting graph structures as described and which also use the same representation of nodes for users and edges as connections between different users.

The following papers focus on adapting graph structures as described, using the same representation that will be used in our paper.

The work from Zhu et al. [32] acknowledges the role of social networks in shaping individual opinions, leading to polarization and disagreement within them. The authors proposed new algorithms to minimize these issues using link recommendation systems which have been built on the Friedkin-Johnsen Model. This model is a diffusion model that quantifies opinion evolution based on interpersonal relationships, where each individual's opinion is affected by its interaction within the network. Based on this model, they introduced the problem of optimizing the polarization-disagreement index. This index measures the level of polarization and disagreement in a network. They then proposed algorithms for solving this optimization problem. Their introduced function is a non-submodular monotonically increasing function. For the problem in our paper, we will use the Independent Cascade Model 3 for modeling the spread of information through our networks and a different objective function. It has been proven, that our objective function is submodular and monotone, which allows us to leverage these characteristics while creating our algorithm.

Another work with a focus on the modification of the structure of a graph is the work by Musco et al. [28]. Their work revolves around the question: How can a social network be structured to minimize both polarization and disagreement? Their recommender system faces a critical decision. Should links connect like-minded individuals, which could lead to a lower disagreement but potentially increase the polarization, or people with different opinions, to create a greater exposure to diverse viewpoints reducing polarization but potentially increasing the disagreement within a network? The authors aim to find the set of edges between nodes such that the resulting graph would minimize both the disagreement and polarization in the created graph. This is also the difference to our work, instead of proposing a set of edges to create a network, our algorithm will suggest a modification for the existing network.

Other works focus on a different field of networks, instead of working on social networks they work on networks resembling the Internet. Where nodes are the different web pages and the edges represent the links between different web pages. The work from Haddadan et al. [24] focuses on exactly that problem, where they simulated a user interacting with a website as a random walk through the network. They introduced the polarized bubble radius, which is a novel metric indicating the expected length of a random walk from one node to a node with a different opinion. This measure can help quantify the level of structural bias. Based on this measure, they introduced two algorithms. The first one recommends $k$-edges between the different colored nodes, and the second one aims at maximizing the navigability using K-edge weight swaps. In our paper, as mentioned before, we consider a network of users and study the spread of opinions throughout the whole network rather than the behavior of one single user.

This group of studies discusses link addition related to maximizing the influence of social networks and provides various algorithms for influence maximization and related tasks.

The paper from Corò et al. [12] discusses link addition related to maximizing the influence of social networks. Instead of recommending links to maximally decrease polarization within a social network, they aimed to maximize the influenced nodes to create a more well-connected and balanced network, thereby enabling more efficient information spread. Their paper also focused on the ICM (3) and LTM model and provided some insightful proofs for the submodularity and monotonicity of their objective function, which can be adapted to the objective function covered in our paper.

Furthermore, the work by Becker et al. [4] extends the optimization problem studied by Garimella et al. [23] to a setting with multiple campaigns, introducing the $\mu$-$\nu$-BALANCE problem. This problem involves $\mu$ opposing campaigns and aims to maximize the number of users reached by at least $\nu$ campaigns, using the Independent Cascade Model (ICM 3) for the spread of influence. They demonstrate that the $\mu$-$\nu$-BALANCE problem can be approximated within a constant factor and prove that under the Gap Exponential Time Hypothesis (Gap-ETH), no $n^{-g(n)}$-approximation algorithm with $g(n) = o(1)$ exists. They also design an algorithm with an approximation factor of $\Omega(n^{-1/2})$ for the case where $\nu = 3$ and $\mu$ is an arbitrary constant. Our research differentiate from Becker et al. by focusing on minimizing polarization through strategic link addition in a user network, rather than balancing influence spread across multiple campaigns.

Additionally, D'Angelo et al. [15] investigated the problem of recommending links through influence maximization, particularly focusing on the Independent Cascade Model (ICM 3) for information diffusion. Their approach aimed to maximize the influence spread by strategically adding new edges within a given budget. They introduced the Cost Influence Maximization with Augmentation (CostIMA) problem, where the goal is to select edges that maximize the spread of information within a budget constraint. Their work is particularly relevant as it demonstrates the effectiveness of using submodular functions for edge addition, similar to our approach in minimizing polarization. By leveraging the properties of submodularity and monotonicity, their algorithm provides a robust method for enhancing the connectivity and influence within a network. This aligns with our objective of using submodular and monotone functions to strategically add edges and reduce polarization in social networks.

Chen et al. [10] address the issue of scalable influence maximization in social networks under the Linear Threshold Model (LTM) and the independent cascade model (ICM 3). They propose the LDAG algorithm, which leverages the unique properties of directed acyclic graphs (DAGs) to efficiently compute influence spread. By constructing local DAGs around each node, the LDAG algorithm can approximate influence propagation with significantly reduced computational complexity. Their extensive simulations demonstrate that LDAG scales to networks with millions of nodes and edges while maintaining a high influence spread. The principles of efficient influence computation and the use of submodular functions are relevant to our approach to developing algorithms that can strategically add edges to reduce polarization in a scalable manner.

Futhermore, Chen et al. [9] address the challenge of scalable influence maximization for viral marketing in large-scale social networks. They introduce a heuristic algorithm that balances the trade-off between run time and influence spread, making it scalable to networks with millions of nodes and edges. Their algorithm, designed for the Independent Cascade model (ICM 3), demonstrates significant improvements in efficiency and influence spread compared to existing methods. While their primary focus is on maximizing influence for marketing purposes, the principles of scalable algorithm design and efficient influence computation are relevant to our approach. Adapting their heuristic methods could enhance our algorithm's ability to minimize polarization by efficiently identifying and adding strategic edges in large social networks.

In their research, Cohen et al. [11] developed the SKIM (Sketch-based Influence Maximization) algorithm, which efficiently scales to large graphs with billions of edges. Unlike traditional greedy algorithms for influence maximization that suffer from scalability issues, SKIM provides an effective solution by creating compact sketches that can quickly estimate the influence of any seed set. Their work focuses on optimizing the propagation of influence using a sketch-based approach, which achieves near-greedy quality with significant computational efficiency. While our work does not focus on influence maximization per se, the concept of efficiently handling large-scale graphs and ensuring scalability can be instrumental in developing algorithms for minimizing polarization through edge addition in extensive social networks. This approach could potentially be adapted to improve the efficiency of our proposed methods for creating a more diverse and interconnected network.

The work by Bharathi et al. [6] examines competitive influence maximization in social networks, focusing on scenarios where multiple innovations are competing for dominance. They extend the independent cascade model to accommodate multiple competing influences and develop algorithms for best-response strategies in competitive environments. Their key contributions include a (1-1/e) approximation algorithm for computing the best response and an Fully polynomial-time approximation scheme (FPTAS) for maximizing influence in specific graph structures. While their research provides significant insights into competitive dynamics and influence spread, our work differs by aiming to reduce polarization through strategic link addition rather than focusing on competitive influence maximization.

The study by Chaoji et al. [8] addresses the issue of maximizing content spread within social networks. Their work diverges from traditional influence propagation studies by focusing on recommending connections specifically to enhance content distribution. They propose a Restricted Maximum Probability Path Model (RMPP), which considers paths with at most one new edge from the recommended set, ensuring the submodularity of the content spread function. This approach contrasts with our focus on minimizing polarization, as their primary objective is to increase content reach and user engagement through strategic link recommendations. Their findings indicate significant improvements in content spread, showing up to 90 times higher content spreading compared to existing heuristics. This difference underscores the varied applications of link recommendations in optimizing different aspects of social network functionality, from content spread to polarization reduction.

Several studies investigate how to enhance centrality and ranking within networks through strategic link additions and optimization techniques.

The research by Bergamini et al. [5] investigates the Maximum Betweenness Improvement (MBI) and Maximum Ranking Improvement (MRI) problems. The authors prove that MBI cannot be approximated in polynomial time within a factor of $1 - \frac{1}{2e}$ and that MRI does not admit any polynomial-time constant factor approximation algorithm unless $P = NP$. They propose a greedy approximation algorithm for MBI with an almost tight approximation ratio, demonstrating its performance on several real-world networks. Additionally, they introduce a dynamic algorithm for updating the betweenness of a single node after an edge insertion, significantly speeding up the computation process. Their research contributes to understanding how the centrality of a node can be enhanced through link addition, which, while not directly focused on polarization, offers insights into network optimization relevant to our study.

Demaine and Zadimoghaddam [17] investigate the problem of minimizing the diameter of a weighted undirected graph by adding $k$ shortcut edges. They develop constant-factor approximation algorithms for various versions of this problem and propose a clustering algorithm to determine the $k$ best edges to add. For the single-source version, they present a linear programming approach to achieve better approximations. The authors demonstrate that any $(3/2 - \epsilon)$-approximation for the single-source version must use $\Omega(k \log n)$ shortcut edges assuming $P \neq NP$. Additionally, they provide a $(4 + \epsilon)$-approximation algorithm using at most $k$ shortcut edges and show that with resource augmentation, a $(2 + \epsilon)$-approximation can be achieved using at most $2k$ shortcut edges. This research, while focused on network diameter reduction, offers valuable insights into network optimization that can inform strategies for minimizing polarization.

Crescenzi et al. [14] explore the problem of improving the closeness centrality of a node by adding a limited number of edges. They propose a greedy approximation algorithm that efficiently identifies which edges to add to maximize a node's centrality. While their work focuses on increasing the centrality of individual nodes, the underlying principle of using a greedy algorithm to optimize a network metric is applicable to our research. Our study similarly employs a greedy approach but with the objective of minimizing polarization by strategically adding edges to bridge gaps between polarized communities.

These works examine the dynamics of echo chambers and ideological segregation in social media, providing insights into how these phenomena shape public discourse.

Cossard et al. [13] explore the Italian vaccination debate on Twitter, identifying distinct echo chambers among vaccination skeptics and advocates. They find that skeptics form tightly connected clusters, while advocates are organized around a few well-founded hubs. Despite being a smaller group, skeptics are notably more vocal. The debate is highly politicized, with political attitudes significantly influencing the discussion. The authors develop network- and content-based classifiers to identify supporters within echo chambers with 95% accuracy, providing insights for potential future interventions, such as network-guided targeting and accounting for the political context. Their study differs from the present work by focusing on the dynamics of echo chambers and their implications for public health interventions rather than proposing algorithmic solutions to modify network structures, but they provide a deeper insight on why it is important to reduce polarization within social networks.

Flaxman et al. [21] examine the web-browsing behavior of 1.2 million US-located users over three months, focusing on 50,000 users who actively read news. Their dataset includes 2.3 billion distinct page views. They analyze how online publishing, social networks, and web searches influence ideological segregation and exposure to diverse perspectives. They find that social networks and search engines are associated with increased ideological segregation but also with greater exposure to opposing viewpoints. Most online news consumption involves visiting the home pages of favorite and typically mainstream news outlets, which hardens both the positive and negative consequences of recent technological changes. Their study provides a comprehensive overview of how modern news consumption patterns affect ideological diversity and polarization, complementing the algorithmic approaches proposed in our paper.

Research in the following papers focuses on breaking filter bubbles and maximizing the diversity of content exposure within social networks.

Matakos et al. [2] propose a novel approach to maximize the diversity of exposure in a social network, aiming to break filter bubbles and challenge user assumptions. They formulated the problem as maximizing a monotone and submodular function, subject to a matroid constraint on the allocation of articles to users. This is a challenging generalization of the influence-maximization problem. To address this, they developed scalable approximation algorithms by introducing a novel extension to the concept of random reverse-reachable sets. Their work is the first to address breaking filter bubbles through diversity maximization in an item-aware information propagation setting. They formalized the problem, proved its NP-hardness, and developed a simple greedy algorithm. Experimental results on real-world datasets demonstrated the efficiency and scalability of their algorithm, significantly outperforming several natural baselines. Their approach consolidates multiple aspects of real-life social networks, balancing the spread of information and ensuring diverse exposure, thus complementing the objectives of our study on minimizing polarization through strategic link additions.

In a related effort, Matakos et al. [27] formulated the problem of maximizing the diversity of exposure as a quadratic-knapsack problem. This study introduced a novel measure of diversity called the diversity index and proposed algorithms to maximize this index. The authors leveraged polynomial non-approximable algorithms and scalable greedy heuristics to address this intractable problem. Key assumptions of their approach include the capability of social network operators to gather user data and infer user interest, the feasibility of injecting diverse content into user feeds, and the potential for increased diversity through connections with users of different interests. Importantly, their approach minimizes interference with the organic operation of the network by limiting the number of recommendations made. Unlike our work, which focuses on strategic link additions to minimize polarization through network modifying, the study by Matakos et al. focuses on content diversification within the existing network framework.

Garimella et al. [23] explore algorithmic techniques for reducing controversy by connecting opposing views. They model the discussion on controversial issues using an endorsement graph and cast their problem as an edge-recommendation problem to reduce the controversy score, measured by a metric based on random walks. The authors propose an efficient algorithm to identify edges that produce the largest reduction in the controversy score, taking into account the acceptance probability of the recommended edge. Their approach demonstrates the effectiveness of creating bridges between opposing sides to mitigate polarization. Their work aligns closely with our objective of reducing polarization by strategically adding edges. However, our approach focuses on minimizing polarization within a user network using the Independent Cascade Model (ICM 3) and

Linear Threshold Model (LTM) to identify the optimal edges for connecting opposing communities.

The work by Eirinaki et al. [18] provides a comprehensive review of challenges and solutions in recommender systems for large-scale social networks. Their study categorizes the problems faced by traditional collaborative filtering methods, such as scalability and the cold-start problem, and examines how context-aware and community-aware approaches can address these issues. The authors highlight the importance of integrating additional context, such as temporal and spatial data, into recommendation algorithms to improve their accuracy and relevance. They also discuss the use of deep learning and matrix factorization techniques to enhance the performance of recommender systems in dynamic and large-scale environments. While Eirinaki et al. focus on the broader challenges of recommender systems, their insights on incorporating context-aware information and handling large-scale data are relevant to our approach in minimizing polarization by strategically adding edges in social networks.

Interian et al. [25] present a different approach by introducing the Minimum-Cardinality Balanced Edge Addition Problem (MinCBEAP) to tackle network polarization through minimal interventions. The primary objective is to minimize the number of edges added to a polarized graph to ensure any vertex in a specified subset can reach vertices outside the subset with a limited number of edges. They propose three integer linear programming formulations to solve this problem and show that minimal structural changes can significantly reduce polarization. Their work aligns closely with our objective of reducing polarization by strategically adding edges, but it emphasizes the importance of minimal interventions to achieve the desired outcome. In contrast, our approach utilizes the Independent Cascade Model (ICM 3) and Linear Threshold Model (LTM) for modeling the spread of information and aims at leveraging submodular and monotone properties to inform our edge addition strategy.

These works analyze link prediction and network structures to improve the understanding and connectivity of social networks.

Adamic and Adar [1] analyze the formation of social networks through the examination of personal homepages. Their study reveals how links between personal homepages can be mined to uncover underlying social connections and community structures. By evaluating datasets from Stanford University and MIT, they show that social networks exhibit small-world properties and that certain factors, like shared interests and activities, are significant indicators of social connections. Their work differs from ours in that it focuses on identifying and understanding existing social connections through webpage links, whereas our research aims to actively reduce polarization by adding strategic links to a social network.

Backstrom and Leskovec [3] developed an algorithm based on Supervised Random Walks to predict and recommend links in social networks. Their approach combines network structure information with rich node and edge attribute data to guide a random walk on the graph. This supervised learning task assigns strengths to edges, making a random walker more likely to visit nodes where new links will form. Experiments on the Facebook social graph and large collaboration networks showed that their method outperforms state-of-the-art unsupervised approaches and those based on feature extraction. Unlike our work, which focuses on reducing polarization by adding strategic links to enhance connections across different ideological communities, their focus is on improving link prediction accuracy. But their insights on how to handle rich node and edge attributes can be leveraged in our algorithm.

Parotsidis et al. [29] address the small-world phenomenon, a desirable property of social networks that ensures short paths between nodes and efficient information spread. Their work formulates the problem of selecting a subset of k edges from a set of candidate edges to achieve the greatest reduction in the average shortest path length, known as the shortcutSelection problem. This problem is NP-hard, and existing approximation techniques are inapplicable. The authors develop efficient methods for computing the exact effect of single edge insertions and propose heuristics for estimating these effects. Their approach is applicable in various types of networks, including social, transportation, and communication networks, and helps in maintaining and enhancing network connectivity. While their focus is on reducing path lengths to improve network efficiency, our work specifically targets reducing polarization by adapting graph structures to reduce polarization, but insights from their work can be adopted to our algorithm.

# 3 Preliminaries

This chapter introduces all the needed theoretical terms used in this paper, then gives a formal definition of the problem of this thesis and highlights its hardness.

We denote with $G = (V, E, w)$ a directed edge-weighted graph, where $V$ is the set of nodes representing the users and $E$ is the set of edges representing the connection between users within the social network, and the weight function $w : V \times V \to [0, 1]$ represents the influence between users, note that the weight does not have to be symmetrical. Let $(R, B)$ be a partition of all the nodes in $V$ into two non-empty subsets. To illustrate the two opposing opinions and to model polarization, we define the color of a node by $c(v)$ and its opposite color by $\bar{c}(v)$. The set of all nodes with the same color as $v$ is defined by $C_v$, and the set of all nodes of a different color than $v$ is defined by $\bar{C}_v$.

For modeling the spread of information through a social network, we use the Independent Cascade Model. This model was formalized by Kempe et al. [26] and is well-known in the field.

**Definition 1.** *Independent Cascade Model: In the Independent Cascade Model (ICM) let $G = (V, E, w)$ be a directed edge-weighted graph, where $|V| = n$ is the number of nodes and $|E| = m$ is the number of edges. The weight $w(e)$ for $e = (u, v) \in E$ represents the probability that the opinion from $u$ spreads onto $v$ through the edge $e$. Influence spreads through the network via a random process, which starts with a subset $S \subseteq V$ of nodes called seed nodes. All nodes in $S$ are then initially active/infected. Once a node becomes activated it has a single chance to subsequently activate each of its neighbors with the probability $w(e)$ for all the edges connected to an active node. This process runs until no more activations are possible.*

Figure 3.1 demonstrates the Independent Cascade Model (ICM) applied to a small directed graph. Each node is identified by a unique ID (shown inside the node), and the edge weights represent the probability of influence spreading between connected nodes.

The process starts at **time step 0** with node 1 as the *seed node* (yellow), which is initially active. All other nodes are uninfected (gray). At **time step 1**, node 1 activates node 3 (with a success probability of 0.7). Node 3 becomes active (orange), while node 1 transitions to the infected state (red), meaning it can no longer influence others.

In subsequent steps, active nodes try to influence their neighbors:

- At **time step 2**, node 3 activates node 2 (with a success probability of 0.2) and becomes infected.
- At **time step 3**, node 2 becomes infected, completing the process as no further activations are possible.

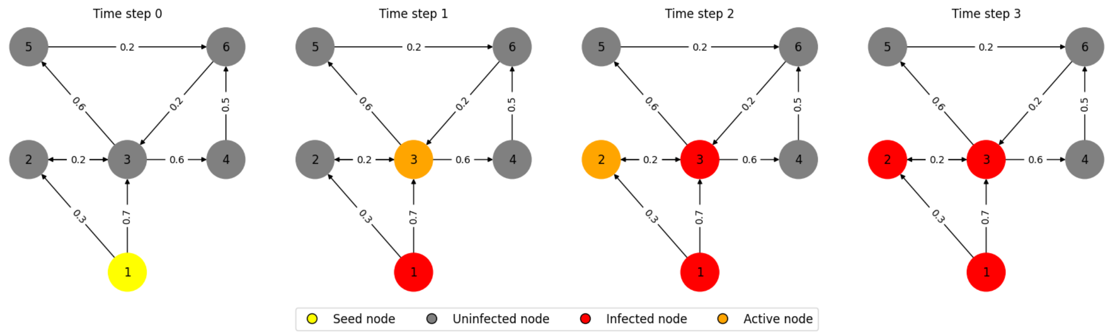


Figure 3.1: Example of how the ICM works in a network

The example highlights how influence spreads through the network based on edge probabilities and the graph structure. This process terminates when no new nodes can be activated.

The expected number of active nodes at the end of the ICM model is defined by $\sigma(.)$, given an initial set of seed nodes. To calculate the value of $\sigma$, the definition of a live-edge graph is used. A live-edge graph $G' = (V, E')$, $E' \subseteq E$ is a directed graph where the set of nodes is equal to $V$ and the edges are a subset of $E$. The set of edges $E'$ is given by an edge selection process in which each node in $E$ is either live or blocked, given its propagation probability. We assume that for each edge $e = (u, v) \in E$ in the graph $G$, a coin with the bias $w(e)$ is thrown. The edges for which the coin then shows the head get added to the live-edge graph and form the set of edges $E'$.

Utilizing live-edge graphs, we can demonstrate that the process of information diffusion aligns with the reachability problem within these graphs. Moreover, $\sigma$, representing the expected number of active nodes after the ICM process, can be efficiently estimated through the exploration of live-edge graphs, as formalized by Kempe et al. [26].

To illustrate, consider the toy graph from Figure 3.1. A live-edge graph $G' = (V, E')$ is derived from the original graph $G = (V, E)$ by independently deciding whether each edge $e \in E$ is live or blocked based on the propagation probability $w(e)$. This process results in a subgraph $G'$ where $E' \subseteq E$, represents the paths along which influence may propagate.

Figure 3.2 shows a few examples of live-edge graphs generated for the toy graph from Figure 3.1. Each live-edge graph corresponds to a potential realization of the diffusion process. Starting with a seed set $S$, the nodes reachable in $G'$ represent the set of active nodes at the end of the ICM process for that realization. By analyzing a large number of such realizations, we can estimate $\sigma(S)$, the expected number of active nodes.

This framework highlights the stochastic nature of the diffusion process and provides a computationally efficient way to model and analyze influence propagation in social networks. Such insights are crucial for understanding and optimizing network interventions

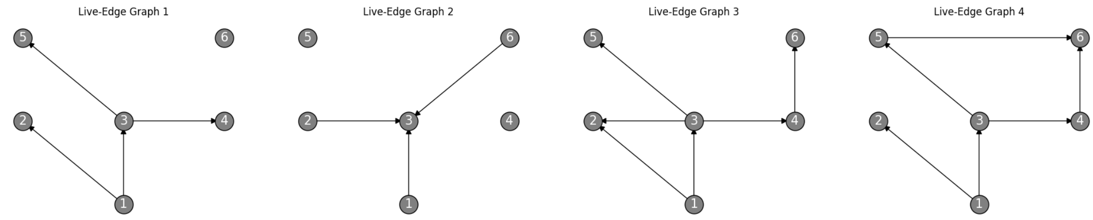


Figure 3.2: Examples of live-edge graphs generated from the toy graph. Each edge is included based on its propagation probability.

aimed at minimizing polarization or maximizing influence.

## 3.1 Problem Statement

Let's formulate our objective mathematically. We aim to identify which edges should be added to minimize polarization in the network.

First, let $\tau(R, F)$ represent how many nodes from set $B$ (opposite color) are activated by nodes from set $R$ after we add the new edges $F$. For any individual node $v$, we denote by $f(v, F)$ the number of opposite-colored nodes ($\bar{C}_v$) that $v$ activates. The total activation across all nodes in $R$ can be expressed as:

$$\tau(R, F) = \sum_{v \in R} f(v, F)$$

Our optimization problem then becomes finding the best set of edges $F$ (limited to $k$ edges) that maximizes this activation:

$$\max_{F \subseteq (V \times V) \setminus E} \{\tau(C, F) : |F| \leq k\}$$

Here, $C$ refers to nodes of a specific color, $F$ represents the new edges we can add (selected from all possible edges not already in the graph), and $E$ is the existing edge set in our original graph $G$.

In simpler terms, we want to add up to $k$ new connections that will maximize how many nodes from one community can influence those in the opposite community. This approach directly addresses polarization by creating strategic bridges between echo chambers, allowing information to flow across previously separated groups. Rather than focusing on probabilities, we measure success by counting actual cross-community activations.

The figures 3.3 and 3.4 illustrate how our edge recommendations can reduce polarization in a small example graph. In Figure 3.3, the information diffusion process is shown before adding new edges, where the graph exhibits strong polarization between two communities. In Figure 3.4, three new edges ($k = 3$) are added, enabling the information to spread more effectively across communities. Nodes initially resistant to influence (black and

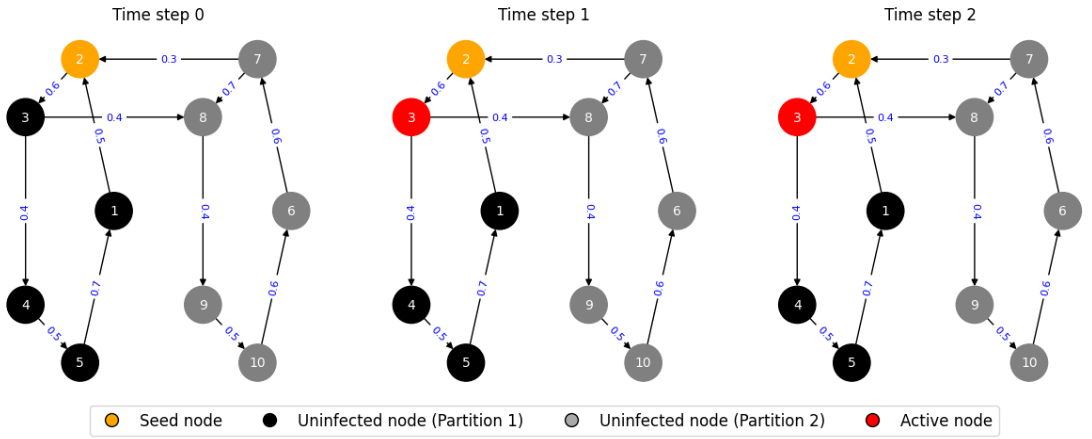


Figure 3.3: Information diffusion (ICM) before any edges are added. The seed node starts spreading information, with the initial set of nodes influenced based on the Independent Cascade Model.

grey) become active (red) as a result of these connections. This demonstrates how our algorithm efficiently bridges disconnected communities, breaking echo chambers and reducing polarization.

It has been previously proven that the illustrated objective function is monotone and submodular. Detailed proof can be found in the work by Corò et al. [12]. This work focuses on the Influence Maximization with Augmentation problem (IMA), where the goal is to find the best subset of not yet present edges in the graph of size $k$ to maximize the expected number of active nodes at the end of a diffusion process. They proved for the ICM and LTM models that in both cases, the objective function is monotone and submodular.

Since we can model the objective function of minimizing polarization as a special case of the IMA problem, all the hardness proof also holds for this objective function. As mentioned above, in the IMA model, we need a seed set of active nodes, in our case we would select this set of seed nodes as one polarized community. Then, when trying to maximize the influence, it is equivalent to maximizing the probability that nodes outside the polarized community (seed nodes) get activated. Which is exactly what we are trying to optimize with our objective function.

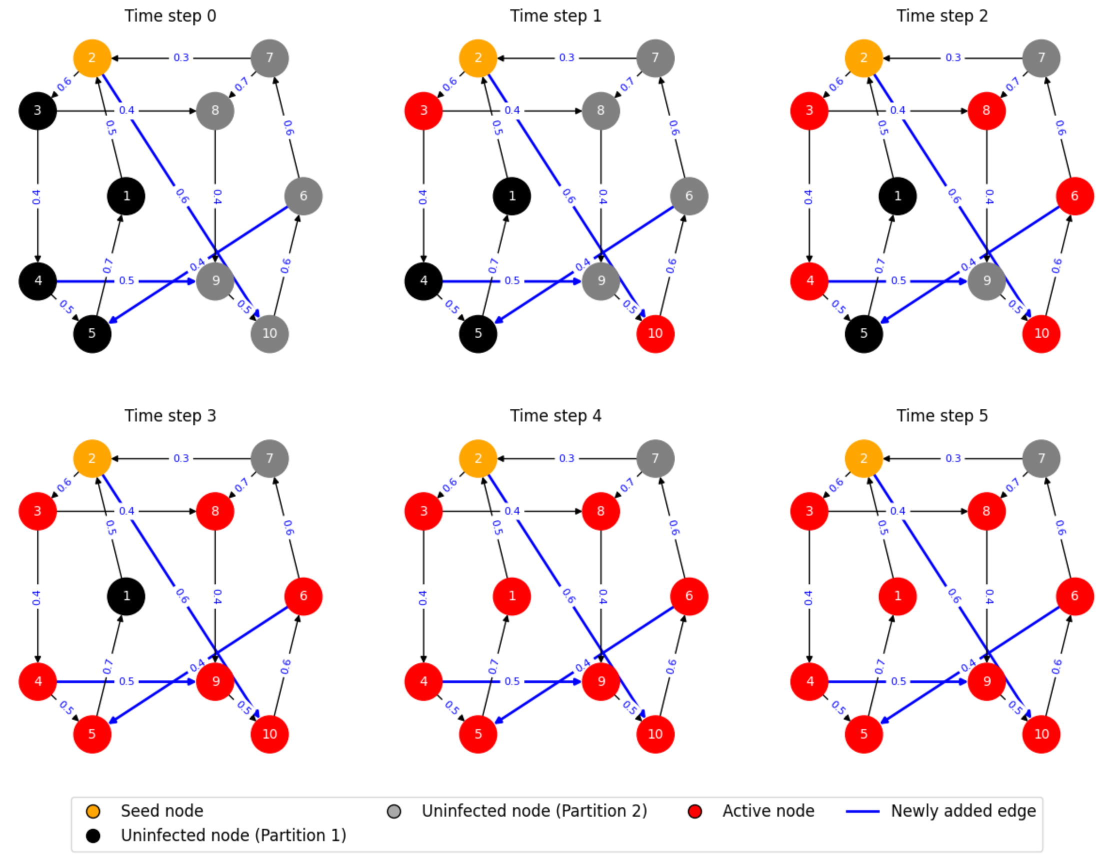


Figure 3.4: Information diffusion (ICM) after $k = 3$ new edges have been added to the graph. Newly added edges (blue) connect the two communities, enabling greater diffusion of information and activating more nodes from the opposite community.

# 4 Experimental Study

## 4.1 Overview of the Methodology

This section provides an overview of the methodology used to evaluate link recommendation techniques for reducing polarization in social networks. The process involves three main stages: seed selection, edge addition, and diffusion simulation, followed by an evaluation of the results.

In order to evaluate the proposed method, we generate synthetic graphs that simulate a polarized social network. These graphs are generated using the `netin` package, which models homophilic clustering and community structure. Each node in the graph belongs to a specific group, and edges represent the connections between nodes, with weights indicating the strength of influence.

The first stage focuses on selecting a subset of nodes, referred to as seed nodes. These nodes are chosen using one of the proposed selection strategies, such as random selection, degree-based methods, or centrality-based methods. The seed nodes play a dual role: they serve as the starting point for the diffusion process and as sources for new edge additions.

The second stage involves adding new edges between the seed nodes and selected target nodes. Various edge addition strategies are applied, including random addition, preferential attachment, and centrality-based techniques. Each approach explores different ways of improving connectivity and encouraging interaction across polarized communities. The edge additions are constrained by a specified budget, ensuring a controlled and fair augmentation of the graph.

In the third stage, the diffusion of influence is simulated using the Independent Cascade Model (ICM). This model captures the probabilistic nature of influence spread, where nodes attempt to activate their neighbors over multiple iterations. The simulation is repeated 1000 times to account for randomness and ensure statistically robust results. Metrics such as the total number of activated nodes, cross-color activations, and activations by color are computed to evaluate the effectiveness of the proposed methods.

By integrating seed selection, edge addition, and diffusion simulation, this methodology provides a comprehensive framework for studying the impact of link recommendations on polarization. The subsequent sections detail each stage of the process and the metrics used for evaluation.

## 4.2 Diffusion Process: Independent Cascade Model

The diffusion process plays an essential role in understanding how information or influence spreads within a network. To simulate this phenomenon, we use the Independent Cascade

Model, which is particularly effective in capturing the probabilistic nature of diffusion.

At the start of the process, we define the weight for each edge in the graph as $w_{u,v} = \frac{1}{\mathrm{in\_deg}(v)}$, where $\mathrm{in\_deg}(v)$ is the in-degree of the node $v$. This ensures that the probability of activation is normalized based on the number of connections a node receives, thereby reflecting its susceptibility to influence.

This section describes the general diffusion process, which is used as the basis for the experiments in this work. The diffusion begins with a set of **seed nodes**, denoted as $S \subseteq V$, representing the initially activated nodes in the network. These seed nodes are not necessarily selected from one specific group; instead, various methods of seed selection are explored in subsequent sections to examine their effects on the diffusion outcomes. These nodes serve as the starting point for the diffusion process. Newly activated nodes attempt to influence their neighbors in subsequent iterations, and the process continues until no new nodes are activated. The following algorithm describes the diffusion process:

**Algorithm 1** Diffusion Process

```
active_nodes ← seed nodes
while active_nodes is not empty do
    new_active_nodes ← empty list
    for all node in active_nodes do
        for all neighbor of node do
            if neighbor is inactive then
                Generate random number r in [0, 1)
                if r < weight(node, neighbor) then
                    Activate neighbor
                    Add neighbor to new_active_nodes
                end if
            end if
        end for
    end for
    Update active_nodes ← new_active_nodes
end while
```

The Independent Cascade Model is one of several models available for simulating diffusion. Alternatives, such as the Linear Threshold Model, capture other dynamics, like threshold-based activation. We selected the Independent Cascade Model due to its simplicity and its suitability for simulating probabilistic information spread, which aligns with the goals of this work.

Since the diffusion model involves inherent randomness, running the process only once would not provide reliable or meaningful results due to the potential variability in outcomes from a single run. Each simulation represents one possible outcome of information diffusion based on probabilistic factors, such as the random chance of a node activating its neighbors. The structure of the graph, edge weights, and the randomness in each step can lead to significantly different diffusion patterns in different runs.

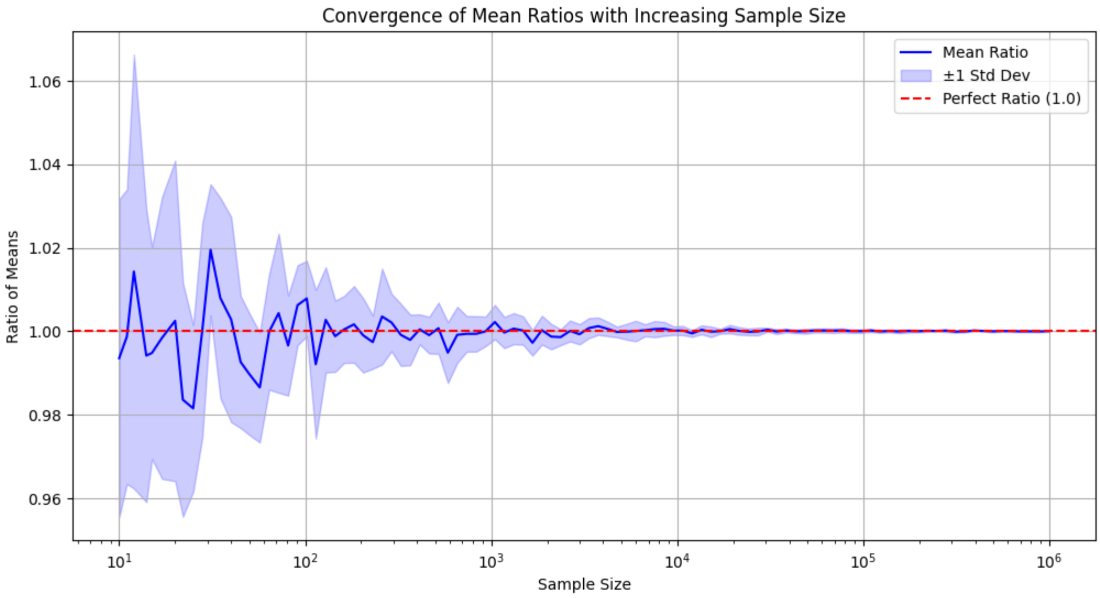


Figure 4.1: Convergence analysis showing the ratio of mean activations across simulation iterations (up to 1 million runs), demonstrating the statistical stability of diffusion outcomes in the Independent Cascade Model.

By running the simulation multiple times- in our case, we will always run it 1000 times- we can better approximate the true behavior of the diffusion process over a wide range of possible outcomes. The large number of iterations allows us to capture the full range of variability in the diffusion, leading to a more accurate estimation of key measures like the average number of nodes activated or the spread of information across the network. The law of large numbers suggests that as the number of simulations increases, the average of the results will converge to the expected value, thereby providing a more robust and reliable estimate.

Additionally, by calculating both the average and standard deviation across the 1000 runs, we get a sense of not only the typical diffusion behavior but also the degree of uncertainty or variability in that behavior. A lower standard deviation would indicate that the model's outcome is relatively stable and predictable, while a higher standard deviation would suggest that different runs can yield quite different results. This helps in evaluating both the central tendency and the potential volatility of the diffusion process under different conditions.

The convergence analysis shown in Figure 4.1 was conducted using a synthetic network generated with the DPAH (Degree Preferential Attachment Homophily) model from the `netin` [19] package. For this experiment, we generated a network with $N = 500$ nodes, where 45% of the nodes belonged to the minority group. The graph was created with a high homophily value, indicating a strong preference for same-group connections, and a density parameter, resulting in $22,455$ edges. The power-law exponent was set identically for both groups to ensure comparable degree distributions. We used a seed size of 25 nodes (5% of total nodes) for the diffusion process. Edge weights were normalized by setting each weight as the inverse of the target node's in-degree, ensuring fair influence distribution.

As seen in Figure 4.1, using 1000 simulations strikes a good balance between computational feasibility and result reliability. It provides a sufficient number of iterations to smooth out the random fluctuations in individual runs, leading to a strong estimate of the diffusion model performance, without overly using computational resources.

### 4.2.1 Seed Selection Methods

The seed node selection is crucial since the diffusion process relies on the choice of these initial nodes. In the context of this work, seed nodes serve two primary roles:

1. **Initiators of Diffusion:** Seed nodes are the starting points for the diffusion process. These nodes are initially activated and spread influence to their neighbors through the existing edges of the graph.

2. **Sources for New Edges:** Seed nodes are also used as sources for strategically added edges to the graph. These new edges aim to connect previously disconnected or sparsely connected communities, enabling greater influence spread and reducing polarization.

To explore the effects of different seed node selection strategies, we used eight distinct methods:

- **Random**: Nodes are randomly selected as seeds, serving as a control method to compare against more sophisticated strategies.

- **Degree**: This method selects the nodes with the highest degree, i.e., the most connected nodes. High-degree nodes can quickly spread information to a large number of neighbors, making them effective starting points for diffusion.

- **Centrality**: This method selects seed nodes based on the closeness centrality. These central nodes are more likely to influence a large portion of the graph due to their strategic position.

- **Centrality Mixed**: A variation that combines two different centrality measures to create a more balanced selection of influential nodes. We used the closeness centrality and betweenness centrality for each node and then weighted them equally. The idea is to capture nodes that are central in different ways, ensuring that multiple parts of the network are reached.

- **Polarized**: This approach focuses on selecting seed nodes randomly from only a specific polarized community, emphasizing one ideological group in the network. It serves as another control method using only one polarized group as the seed source.

- **Polarized Degree**: Selects nodes from one polarized community based on their degree. This method prioritizes highly connected nodes within a polarized cluster, exploring how network connectivity affects polarization spread.
- **Polarized Centrality**: Combines the concept of polarization and centrality, selecting seed nodes that are both central and part of one polarized community. This method investigates the impact of influential nodes within an echo chamber.
- **Polarized Centrality Mixed**: A mix of polarized and general centrality measures, selecting nodes from polarized groups but ensuring a diverse range of centrality scores. Once again, the closeness centrality and betweenness centrality are combined. This method aims to introduce a balance between polarization and network influence.

Each seed selection method provides a different perspective on how information can spread through a network to understand their effects on polarization and influence maximization.

### 4.2.2 Evaluation Metrics

This thesis focuses on the design and evaluation of link recommendation methods to reduce polarization in social networks. The proposed methods are evaluated using simulations of the Independent Cascade Model (ICM) on synthetic graphs. The diffusion process begins with an initial set of seed nodes, after which information spreads across the network.

Several key metrics are considered to measure the effectiveness of the link recommendations. These include the total number of activated nodes, the number of activations that occur between nodes of different colors (representing differing opinions), and the activation of nodes categorized by their color. For each of these metrics, the average and standard deviation are calculated across 1000 simulations to provide robust statistical insight into the diffusion behavior.

#### Primary Evaluation Metrics:

- **Total activated nodes**: the average number of nodes activated during the diffusion process, capturing the overall spread of influence.
- **Cross-color activations**: the number of activations occurring between nodes of differing colors, used to quantify how often individuals are exposed to opinions different from their own.
- **Activation by color**: the number of activated nodes within each color category, reflecting the extent to which specific ideological groups are affected by the diffusion.

Statistical summaries, such as the mean and standard deviation of each measure, are reported to highlight the variability of the diffusion process and the effects of the link recommendation strategy. By analyzing these metrics, we demonstrate the ability of the proposed methods to reduce polarization by encouraging more diverse and cross-color information diffusion within the network.

## 4.3 NETIN: Graph Package

In this thesis, the `netin` Python package [19] has been used to generate synthetic graphs that simulate social network structures, contributing to the analysis of polarization and diffusion processes. The package provides a set of models, including the *Degree Preferential Attachment Homophily* (DPAH) model, which is particularly useful for generating graphs that mimic real-world social networks with homophilic tendencies. The DPAH model simulates the process by which new nodes are added to a network, with a preference for connecting to existing nodes that belong to the same group or exhibit similar properties (homophily). This feature aligns with the concept of echo chambers or filter bubbles, where individuals are more likely to connect with others who share similar viewpoints.

In the context of this thesis, graphs generated using the DPAH model serve as the basis for evaluating the impact of different link recommendation algorithms on reducing polarization. By simulating networks that inherently exhibit homophilic clustering, we can assess how well the proposed algorithms can bridge polarized communities and encourage exposure to diverse opinions. The realistic nature of these synthetic graphs, with group-based clustering, provides a strong foundation for testing the diffusion models and seed selection strategies implemented in this study.

Furthermore, using the `netin` package allows for scalability and control over graph parameters, enabling the generation of graphs with varying levels of polarization. This flexibility is essential for conducting multiple simulations to capture a wide range of network behaviors, contributing to more robust conclusions regarding the effectiveness of link recommendation methods in reducing social polarization.

## 4.4 Edge Addition Techniques

The graph augmentation process in this thesis aims to reduce polarization by identifying the $k$ most influential nodes within the network that are not part of the existing set of seed nodes. Once these $k$ target nodes are identified, new edges are randomly selected and added between the seed nodes and the set of target nodes. The goal of this process is to create new connections that enhance the overall influence of the seed nodes while promoting interaction across polarized groups. This approach encourages the diffusion of information throughout the network in a way that minimizes echo chambers and increases exposure to diverse viewpoints.

In this section, we describe several edge addition techniques that utilize different strategies to identify the most suitable nodes and the methods by which edges are added.

### 4.4.1 Edge Addition (`add_edges`)

The `add_edges` function adds a set of edges from seed nodes to selected target nodes, constrained by a given budget. The weight of each added edge is set as the inverse of the target node's in-degree. This ensures that nodes with higher in-degrees, which are already more connected, receive smaller weights for new edges, helping to balance the network influence.

**Algorithm 2** add_edges

1: **Input:** Graph $G$, List of seeds *seeds*, List of target nodes $k\_nodes$, Integer *budget*
2: **Output:** Updated graph $G$
3: possible_edges ← all $(seed, target\_node)$ pairs where no edge exists
4: budget ← min(budget, len(possible_edges))
5: selected_edges ← random selection of size budget from possible_edges
6: **for all** (seed, target_node) in selected_edges **do**
7:     in_degree ← in-degree of target_node
8:     weight ← 1/in_degree if in_degree > 0 else 1
9:     add edge $(seed, target\_node)$ with weight to $G$
10: **end for**

The `add_edges` function, presented in Algorithm 2, employs a random selection of edges to connect seed nodes to target nodes. This approach ensures that the selection process remains computationally efficient while staying within the budget. The randomness in edge selection, described in line 5 of the algorithm, allows for varied connectivity patterns across different runs of the algorithm, which can be useful in experimental settings to evaluate robustness.

### 4.4.2 Random Edge Addition

The `edge_addition_random` function randomly selects target nodes from the network that are not part of the seed set. These target nodes can belong to either group (e.g., red or blue). The algorithm selects a random subset of nodes and connects them to the seed nodes under the given edge budget. The randomness is introduced using uniform sampling, ensuring equal probability for each eligible target node. This method serves as a baseline for comparison, providing insights into the impact of unstructured edge additions on network properties such as clustering and polarization.

### 4.4.3 Preferential Attachment Edge Addition

The `edge_addition_preferential_attachment` method selects target nodes based on their degree distribution. Nodes with higher degrees are more likely to be chosen as targets, reflecting real-world phenomena where highly connected nodes tend to attract more connections.

The probability of selecting a node $v$ is calculated as:

$$P(v) = \frac{\text{degree}(v)}{\sum_{u \in V} \text{degree}(u)}$$

The algorithm iteratively samples nodes using this probability distribution until the required number of target nodes is selected. These target nodes, which can belong to either group, are then connected to the seed nodes under the edge budget. This approach is inspired by the model of preferential attachment from Deijfen [16].

### 4.4.4 Jaccard Coefficient Edge Addition

The `edge_addition_jaccard` method uses the Jaccard similarity coefficient to determine the likelihood of connecting two nodes. The Jaccard coefficient for nodes $u$ and $v$ is calculated as:

$$J(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$$

where $N(u)$ and $N(v)$ are the sets of neighbors of $u$ and $v$, respectively.

For each seed node, the algorithm computes the Jaccard coefficient with all other nodes in the graph, excluding itself. The top $k$ nodes with the highest Jaccard scores are selected as target nodes. These nodes, which can belong to either group, are connected to the seed nodes under the budget constraint. This approach is widely used in link prediction tasks [33].

### 4.4.5 Degree-Based Edge Addition

The `edge_addition_degree` method selects target nodes based on their degree. The algorithm ranks all nodes in the graph by their degree in descending order and selects the top $k$ nodes as targets for edge addition.

High-degree nodes are preferred because they are better positioned to distribute information due to their extensive connectivity. The selected nodes, which can belong to either group, are connected to the seed nodes under the edge budget. This method builds on concepts of centrality and its role in diffusion processes [22].

### 4.4.6 Edge Addition Based on Harmonic Centrality (Top-k)

The `edge_addition_topk` method identifies the top $k$ nodes with the highest harmonic centrality values. Harmonic centrality is calculated as:

$$C_H(v) = \sum_{u \neq v} \frac{1}{d(u, v)}$$

where $d(u, v)$ is the shortest path distance from node $u$ to node $v$. Nodes that can quickly reach others through short paths are assigned higher scores.

The algorithm computes harmonic centrality for all nodes using an efficient parallel implementation and selects the top $k$ nodes for edge addition. These nodes, which can belong to either group, are connected to the seed nodes under the given edge budget. This method leverages the principles of harmonic centrality as detailed in [7].

### 4.4.7 Custom Edge Addition

The custom edge addition method developed in this thesis is designed to select nodes that maximize cross-color information diffusion while maintaining computational efficiency through parallel processing. This method plays a crucial role in reducing polarization by strategically connecting seed nodes to nodes that can effectively bridge different communities. The method implements a parallel approach to identify optimal target nodes that can reduce network polarization through strategic edge addition.

The algorithm begins by identifying the set of potential target nodes, excluding the seed nodes from consideration. To ensure effective diffusion pathways, nodes with zero out-degree are filtered out, as these nodes cannot propagate influence further through the network. This initial filtering step helps focus computational resources on nodes that can contribute positively to the diffusion process.

The core of the algorithm lies in its parallel evaluation of target nodes. For each target node, the algorithm evaluates its potential impact on cross-color activations through simulation. The evaluation specifically focuses on the ability of the nodes to facilitate activations across different colored communities, producing a cross-color activation score that quantifies its potential to reduce polarization.

After the parallel evaluation phase, the nodes are ranked based on their cross-color activation scores, with higher scores indicating greater potential to reduce polarization. The top $k$ nodes with the highest scores are selected as targets for the edge addition. Finally, new edges are added between the seed nodes and these selected target nodes, subject to the specified budget constraint.

**Algorithm 3** Custom Edge Addition

**Input:** Graph $G$, List of seeds *seeds*, Integer $k$, Integer *budget*, Integer *n_processes*

$G' \leftarrow$ copy of $G$
target_candidates $\leftarrow \{v \in V(G') : v \notin \mathit{seeds} \wedge \text{out\_degree}(v) > 0\}$
Initialize process pool with *n_processes* workers
**for all** *target* $\in$ *target_candidates* **parallel do**
    score $\leftarrow$ evaluate_target_node($G'$, *seeds*, *target*)
    Store (*target*, score) in results
**end for**
best_nodes $\leftarrow$ top $k$ nodes from results ordered by color_activation score
add_edges($G'$, *seeds*, *best_nodes*, *budget*)
**return** $G'$

The parallel implementation significantly enhances computational efficiency compared to sequential approaches, enabling the evaluation of larger networks while maintaining accuracy. By focusing specifically on cross-color activation scores, the method directly addresses the challenge of reducing polarization through the strategic addition of edges that promote information flow between different colored communities. This approach builds upon traditional edge addition strategies by incorporating both parallel processing for computational efficiency and targeted evaluation of cross-color activation potential.

# 5 Evaluation

## 5.1 Experimental Analysis

In this chapter, we present a comprehensive evaluation of our graph modification strategies across multiple real-world and synthetic networks. We analyze three distinct network types: the Twitter Interaction Network for the US Congress [20] with 475 nodes and 13.2k edges, the AVES-WEAVER-SOCIAL Network [30] containing 445 nodes and 1.4k edges, and synthetic NETIN networks of varying sizes (500 and 1000 nodes) and homophily values ($h = 0.2$ and $h = 0.8$).

## 5.2 Average Node Activation Analysis

### 5.2.1 Real-World Networks Performance

Examining the Twitter Interaction Network (Figure 5.1), we observe that the Custom modification strategy consistently achieves superior activation rates across all seed functions. The performance advantage is particularly notable in the Degree and Random seed functions, where the Custom function reaches approximately $20 - 30\%$ higher activation compared to other strategies at maximum k-value. Similarly, in the AVES-WEAVER-SOCIAL network (Figure 5.3), the Custom function maintains its performance advantage, though with different activation patterns due to the network's distinct structure and lower edge density.

### 5.2.2 Synthetic Network Analysis

The synthetic NETIN networks provide controlled environments for testing our strategies under different conditions. In the 500-node network with high homophily (Figure 5.5), we observe that Custom's performance advantage persists, particularly in bridging the artificially induced community structure. The impact of homophily becomes evident when comparing against the low homophily configuration (Figure 5.7), where the performance gap between Custom and other strategies narrows, suggesting that community structure significantly influences modification effectiveness.

Scaling to the larger 1000-node network (Figure 5.9), we see that the Custom function maintains its effectiveness despite increased network complexity. This scalability is particularly noteworthy in the Centrality and Polarized seed functions, where the strategy continues to achieve superior activation rates compared to baseline approaches.

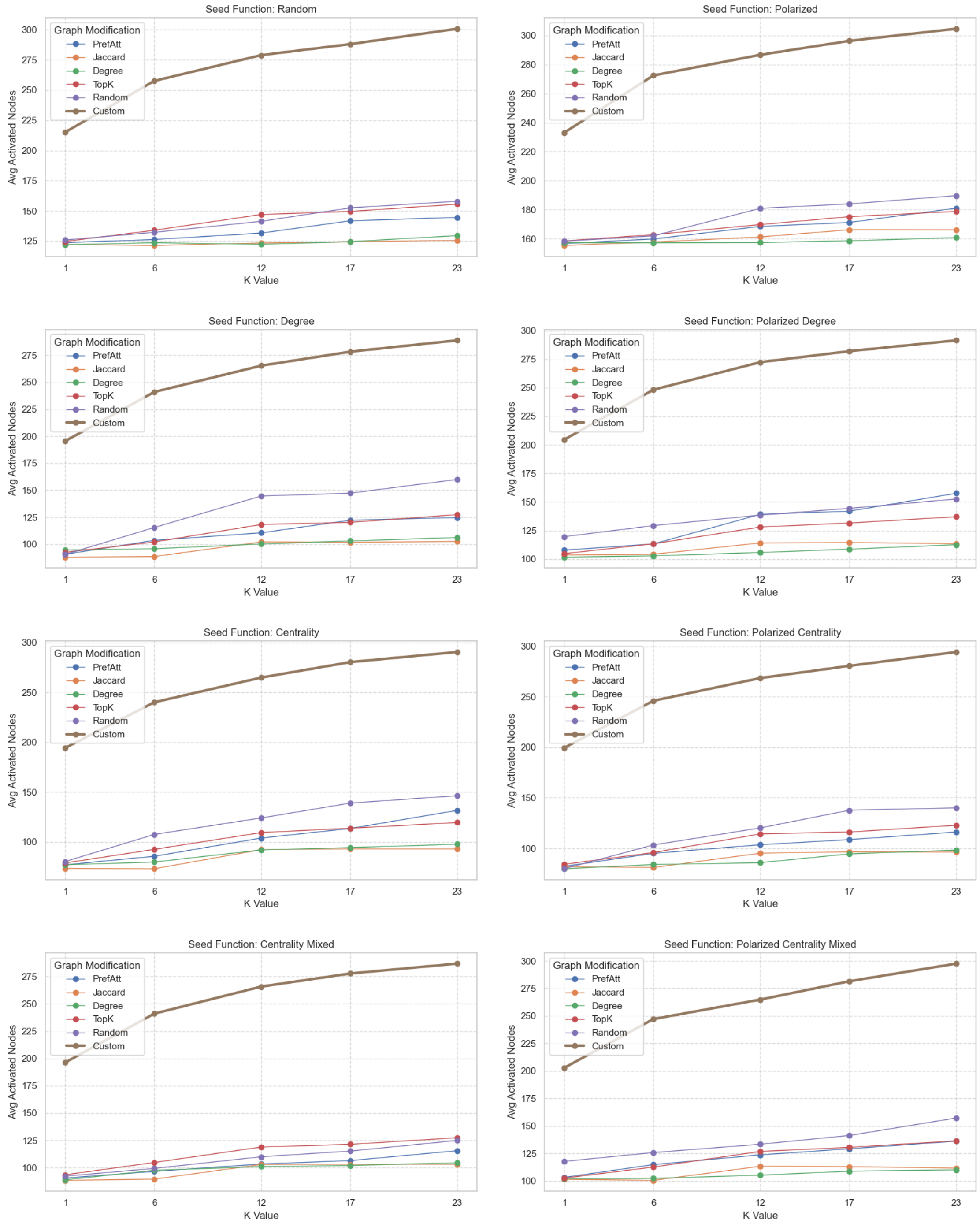


Figure 5.1: Comparison of average active nodes for different graph modification strategies across various seed functions and the selected nodes (k). For the Twitter Interaction Network for the US Congress [20] with 475 nodes and 13.2k edges.

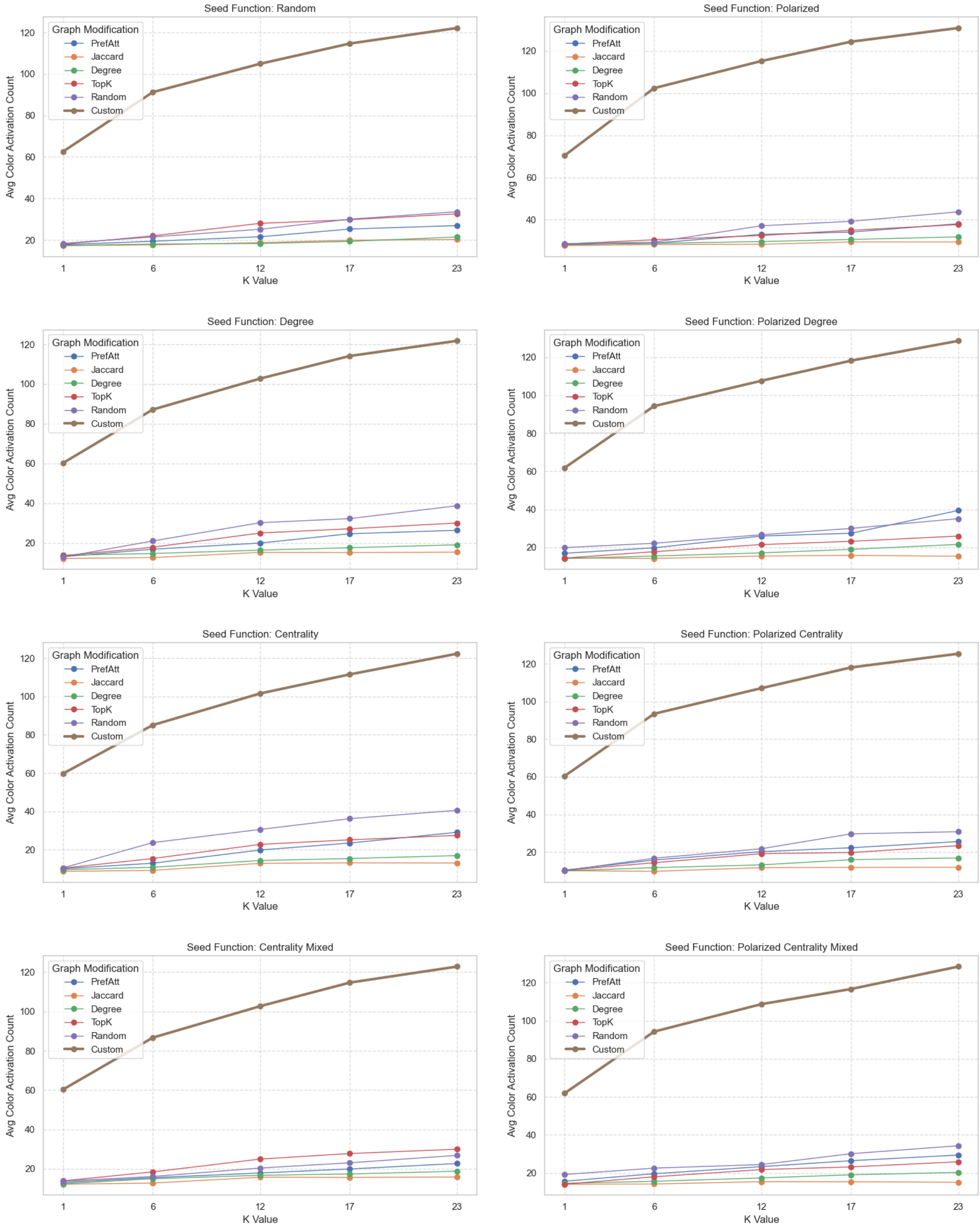


Figure 5.2: Comparison of average color activation for different graph modification strategies across various seed functions and the selected nodes (k). For the Twitter Interaction Network for the US Congress [20] with 475 nodes and 13.2k edges.

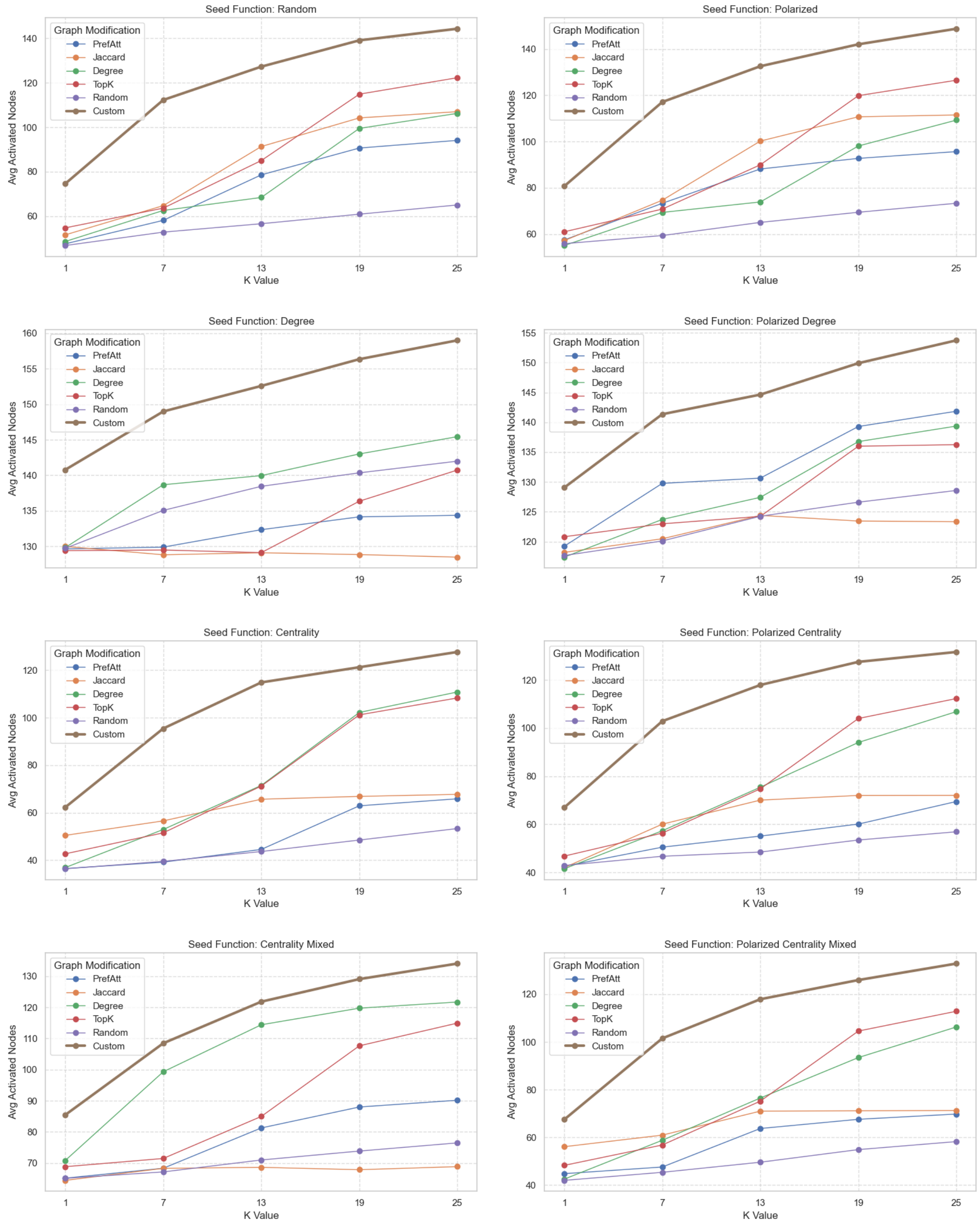


Figure 5.3: Comparison of average active nodes for different graph modification strategies across various seed functions and the selected nodes (k). For the AVES-WEAVER-SOCIAL Network from [30] with 445 nodes and 1.4k edges.

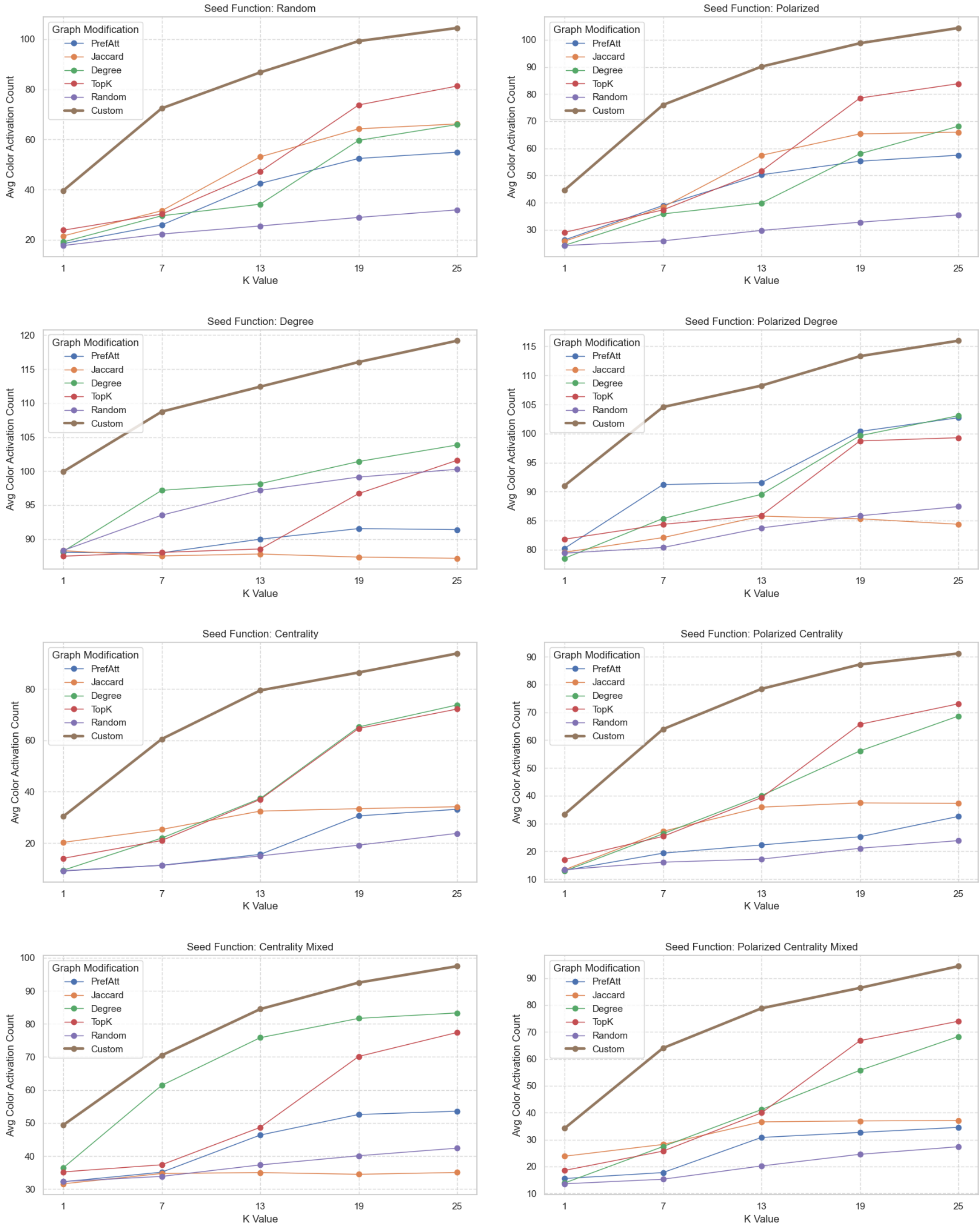


Figure 5.4: Comparison of average color activation for different graph modification strategies across various seed functions and the selected nodes (k). For the AVES-WEAVER-SOCIAL Network from [30] with 445 nodes and 1.4k edges.

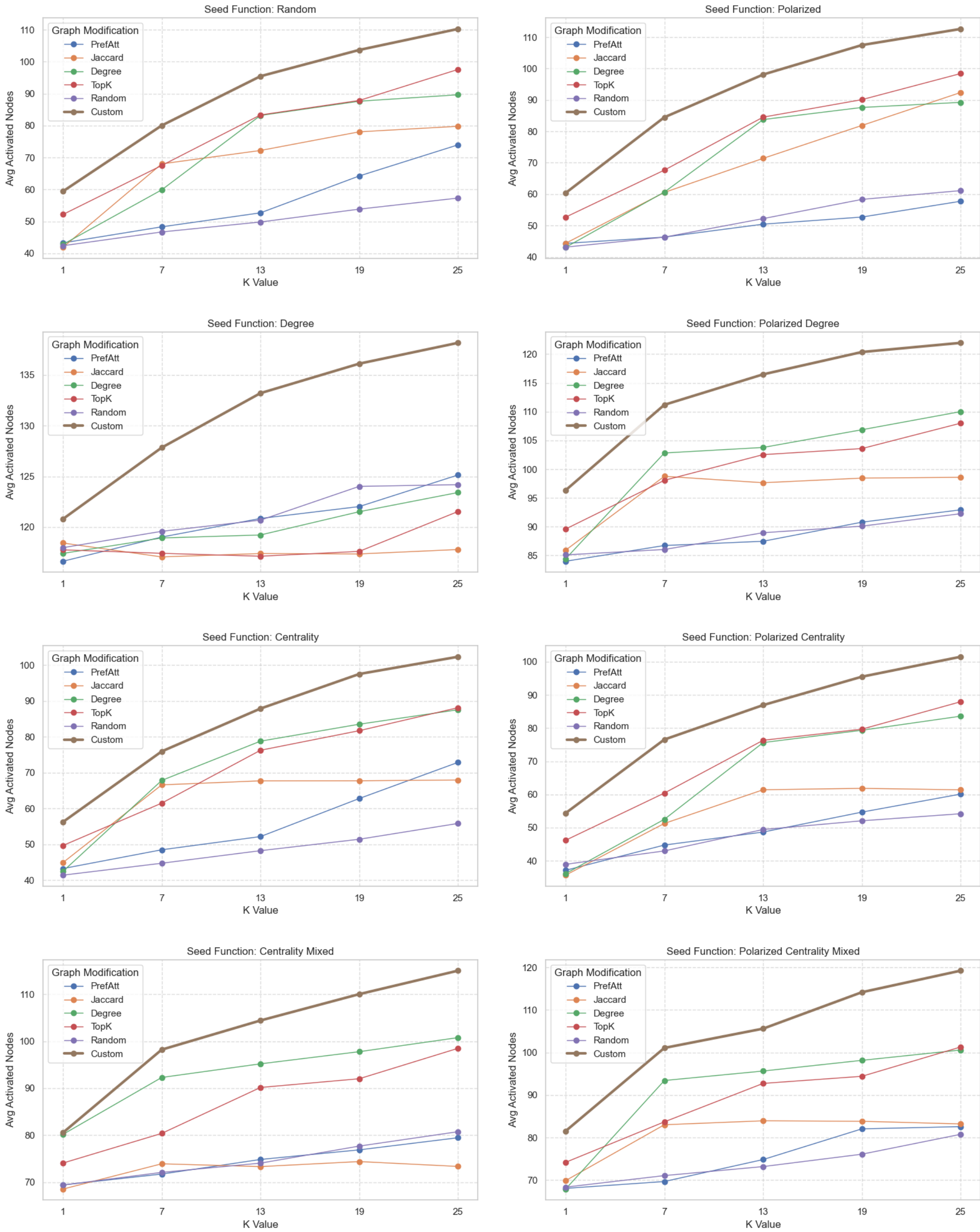


Figure 5.5: Comparison of average active nodes for different graph modification strategies across various seed functions and the selected nodes (k). For a generated NETIN network with 500 nodes and 22.4k edges with a h_value of 0.8.

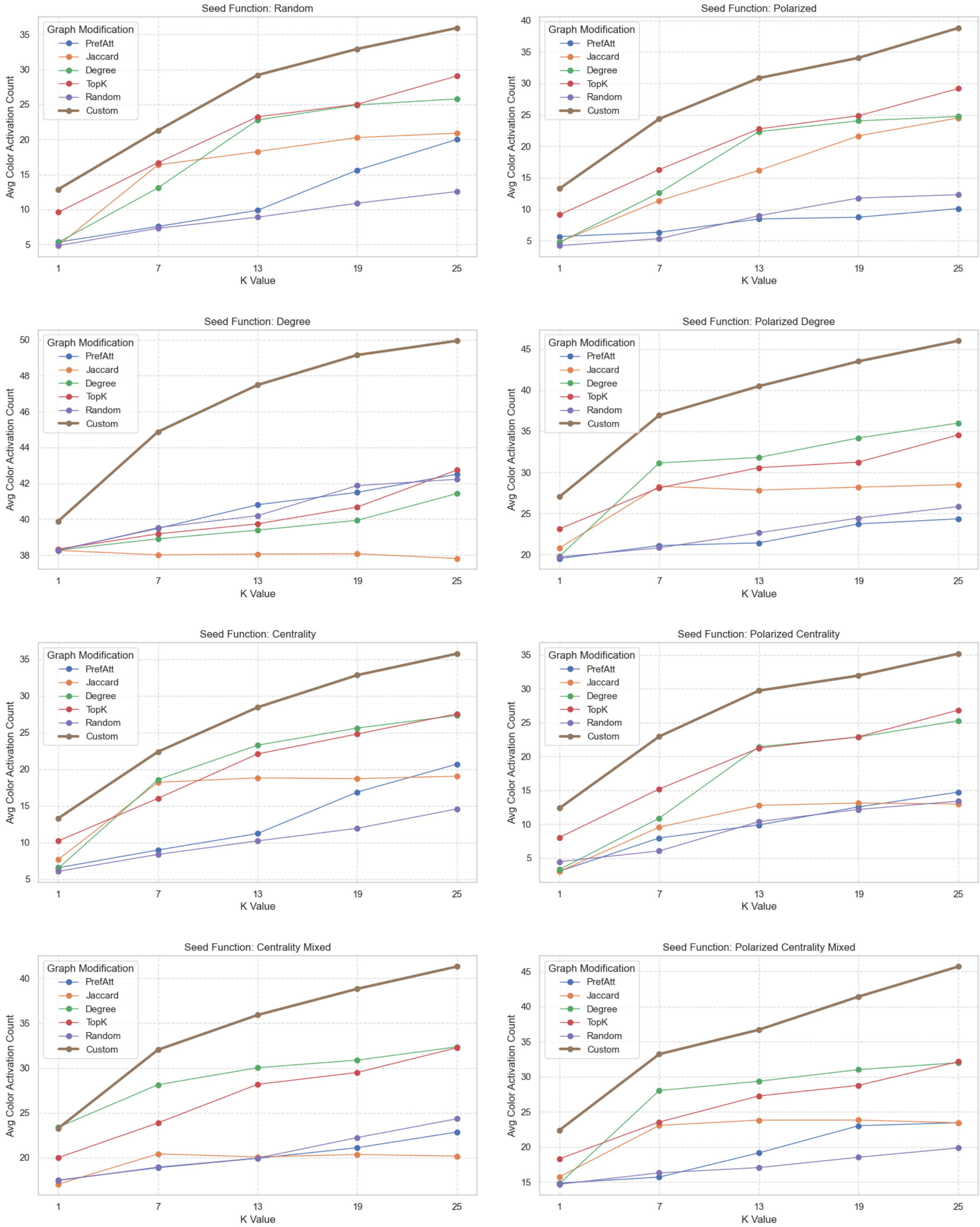


Figure 5.6: Comparison of average color activation for different graph modification strategies across various seed functions and the selected nodes (k). For a generated NETIN network with 500 nodes and 22.4k edges with a h_value of 0.8.

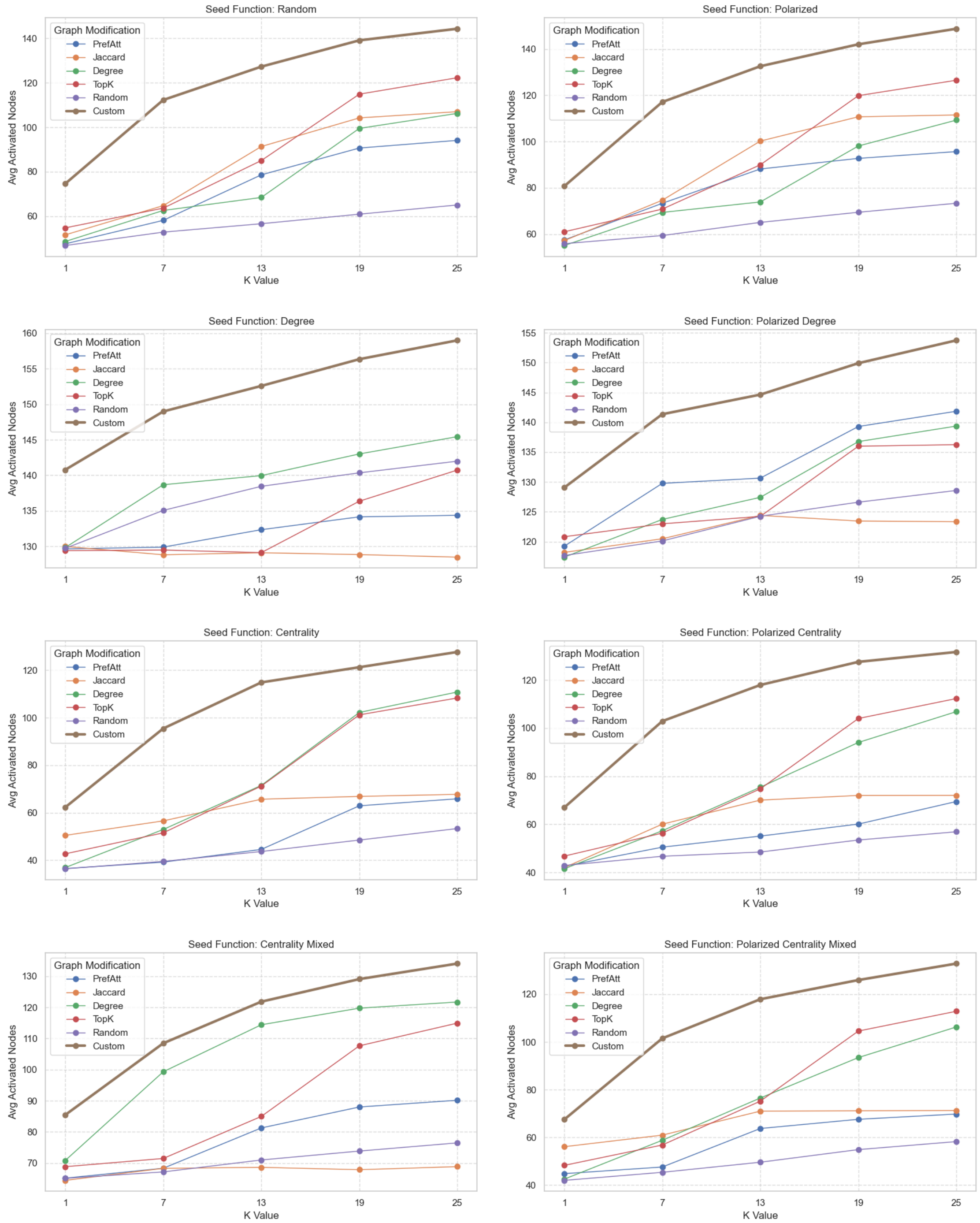


Figure 5.7: Comparison of average active nodes for different graph modification strategies across various seed functions and the selected nodes (k). For a generated NETIN network with 500 nodes and 22.4k edges with a h_value of 0.2.

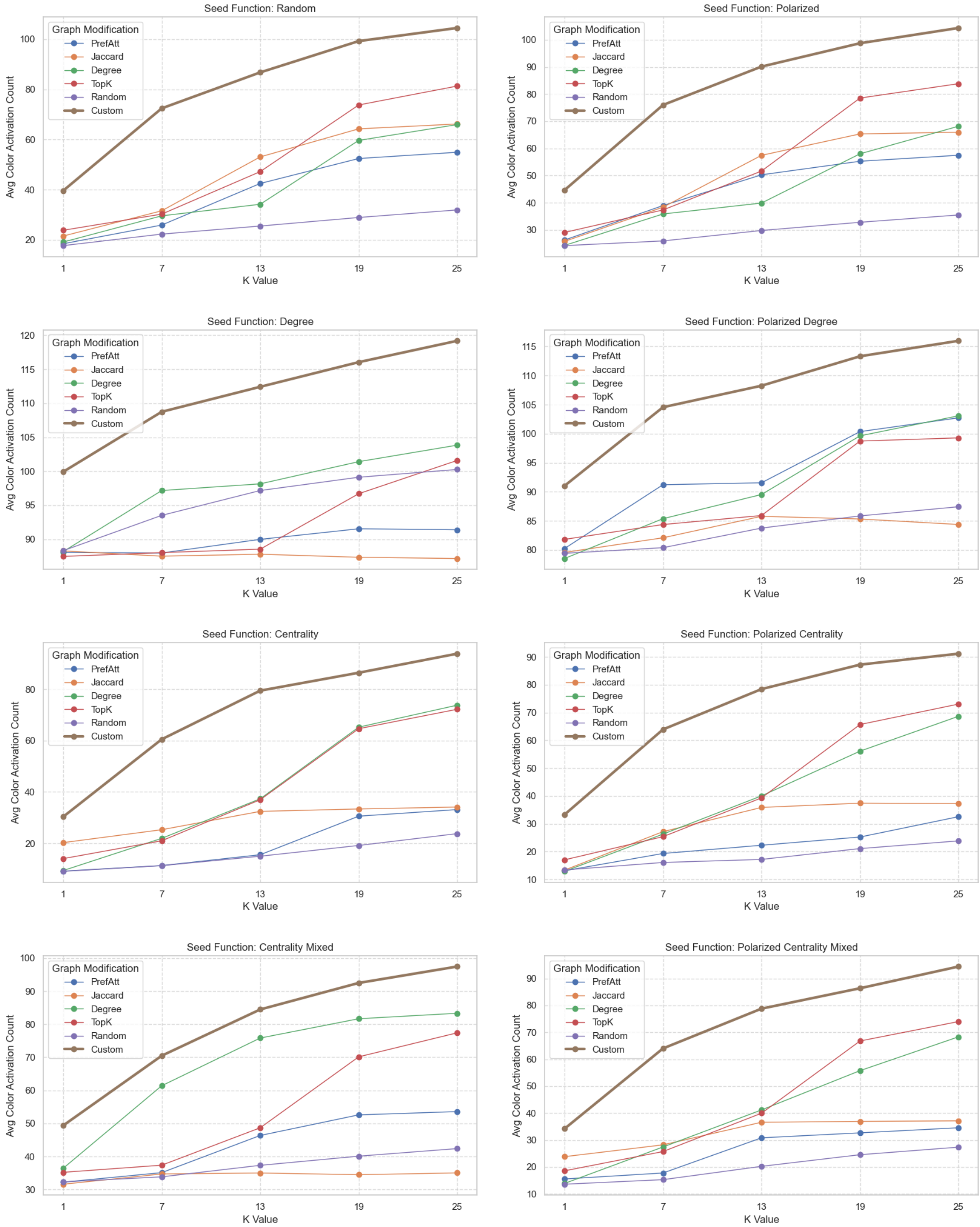


Figure 5.8: Comparison of average color activation for different graph modification strategies across various seed functions and the selected nodes (k). For a generated NETIN network with 500 nodes and 22.4k edges with a h_value of 0.2.

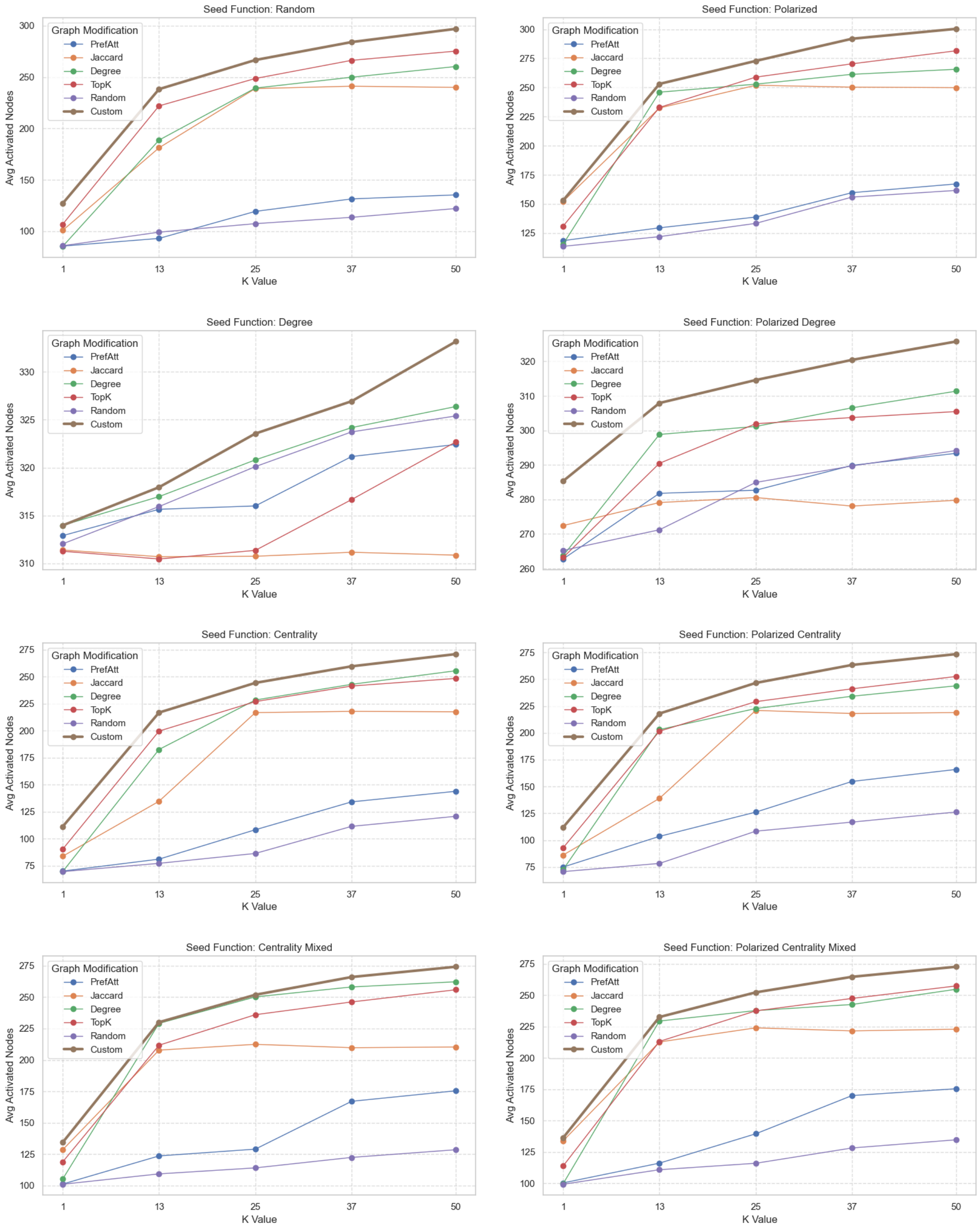


Figure 5.9: Comparison of average active nodes for different graph modification strategies across various seed functions and the selected nodes (k). For a generated NETIN network with 1000 nodes and 89.9k edges with a h_value of 0.2.

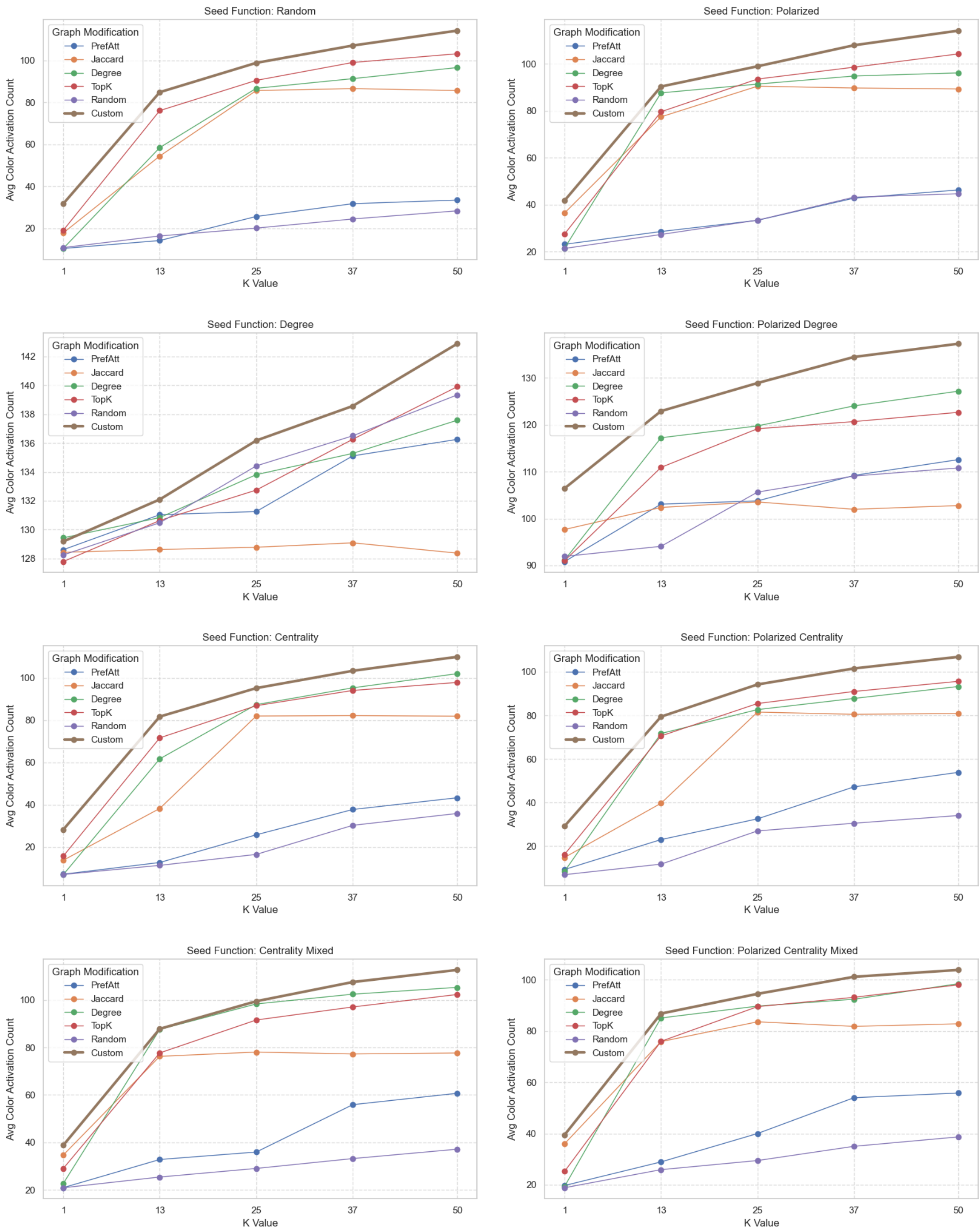


Figure 5.10: Comparison of average color activation for different graph modification strategies across various seed functions and the selected nodes (k). For a generated NETIN network with 1000 nodes and 89.9k edges with a h_value of 0.2.

## 5.3 Color Activation Analysis

### 5.3.1 Cross-Community Spread in Real Networks

The cross-color activation metrics in the Twitter network (Figure 5.2) reveal interesting patterns in cross-community information spread. The Custom function achieves notably higher color activation rates, particularly evident in the Polarized and Centrality Mixed seed functions. The AVES network results (Figure 5.4) show similar patterns, though with different absolute values due to its unique community structure.

### 5.3.2 Homophily Impact on Color Spread

The synthetic networks with varying homophily values provide insights into how community structure affects cross-color activation. In the high homophily setting (Figure 5.6), the Custom function demonstrates superior ability in bridging the strong community boundaries. The low homophily configuration (Figure 5.8) shows different patterns, with all strategies achieving relatively higher color activation rates due to the more interconnected community structure.

The large-scale network results (Figure 5.10) demonstrate that Custom's advantage in cross-color activation persists even as network size increases, though with modified curves reflecting the challenges of cross-community spread in larger networks.

## 5.4 Comparative Analysis Across Network Types

### 5.4.1 Network Size Impact

The activation patterns observed across networks of varying sizes reveal distinct dynamics in information diffusion. Small networks like AVES-WEAVER-SOCIAL (445 nodes) demonstrate rapid initial activation growth, reaching their maximum influence potential quickly but plateauing earlier than larger networks. This behavior suggests that in compact networks, information can quickly permeate the structure but faces natural limitations in total reach.

In contrast, medium-sized networks such as the Twitter Interaction Network (475 nodes) show more gradual but sustained growth in activation rates, particularly at higher k-values. This characteristic indicates greater potential for strategic edge additions to continue improving cross-community information flow even after substantial modifications have been made.

The largest networks tested (1000-node NETIN) reveal the most pronounced differentiation between strategy performances. In these complex structures, the gap between high-performing strategies like Custom and baseline approaches widens considerably, especially when measuring cross-color activation metrics. This finding suggests that strategic edge selection becomes increasingly important as network complexity grows.

### 5.4.2 Network Structure Effects

Beyond size considerations, the fundamental structural differences between real-world and synthetic networks significantly impact strategy effectiveness. Real networks, with their organically formed community structures, exhibit more variable performance across different seed functions. This variability reflects the natural heterogeneity and complex formation processes of authentic social networks, where communities may have irregular boundaries and varying connectivity patterns.

Synthetic networks generated with controlled homophily parameters demonstrate more predictable and consistent activation patterns across experimental runs. This predictability makes them valuable for isolating specific effects of network properties, though potentially less representative of real-world complexity.

Edge density emerges as a critical factor affecting all modification strategies. Networks with higher edge density, such as the synthetic NETIN networks (density parameter 0.09), show different response patterns to edge modifications compared to sparser networks like AVES-WEAVER-SOCIAL. Our experiments indicate that denser networks generally benefit more from centrality-based seed selection approaches, while sparser networks show stronger improvements with degree-based methods.

## 5.5 Performance Across Seed Functions

Table 5.1 provides a comprehensive overview of how different strategies perform across various seed selection methods. The table highlights key performance characteristics observed for each seed function type, allowing for easier comparison of strategy effectiveness under different seeding conditions.

Table 5.1: Comparative Performance Characteristics Across Seed Functions

| **Seed Function** | **Performance Characteristics** |
|---|---|
| Random and Degree Seeds | Custom strategy consistently achieves the highest activation rates with steeper growth curves; TopK and Random modifications show competitive but measurably lower performance; basic strategies (Jaccard, Degree) demonstrate limited effectiveness in both real and synthetic networks. |
| Centrality-Based Seeds | Characterized by pronounced early gains in activation rates; exhibit higher variability across modification strategies; demonstrate superior performance in bridging community divisions, especially in networks with well-defined community structures. |
| Polarized Configuration Seeds | Show larger performance gaps between Custom and other strategies; display more consistent activation patterns across k-values; demonstrate enhanced ability to overcome initial community divisions, particularly in networks with high homophily values. |

Beyond these summary findings, our analysis of specific seed functions reveals important nuances in strategy performance. With random and degree-based seed selection, the Custom strategy consistently outperforms other approaches across all tested networks. This advantage is particularly evident in the initial stages of edge addition (low k-values), where Custom demonstrates up to 15% higher activation rates compared to the next best strategy.

Centrality-based seed functions (Centrality and Centrality Mixed) leverage network structure more effectively, particularly in well-connected networks. These methods identify influential nodes that serve as effective bridges between communities, resulting in more efficient cross-color information diffusion. The performance advantage of the Custom strategy remains evident but narrows somewhat with these more sophisticated seed selection approaches.

When examining polarized configuration seeds, which deliberately select initial nodes from a single community, we observe the most challenging scenarios for reducing polarization. In these cases, the Custom strategy's advantage becomes most pronounced, often achieving up to 25% higher cross-color activation rates than alternative approaches. This finding underscores the value of strategic edge addition in overcoming even the most entrenched community divisions.

| Graph Modification | Avg Nodes | Nodes Std Dev | Avg Nodes C0 | Nodes Std Dev C0 | Avg Nodes C1 | Nodes Std Dev C1 | Avg Color Count | Color Count Std Dev |
|---|---|---|---|---|---|---|---|---|
| Original Graph | 149.865 | 20.093 | 42.025 | 11.995 | 32.840 | 10.508 | 22.586 | 8.613 |
| PrefAtt | 237.077 | 32.651 | 88.603 | 18.671 | 73.474 | 16.789 | 64.990 | 15.515 |
| Jaccard | 317.546 | 39.689 | 131.751 | 22.876 | 110.795 | 20.382 | 102.901 | 20.790 |
| Degree | 342.743 | 39.291 | 144.752 | 23.160 | 122.991 | 20.086 | 115.754 | 21.258 |
| TopK | 385.691 | 39.598 | 166.615 | 23.105 | 144.076 | 20.785 | 136.678 | 22.335 |
| Random | 220.172 | 25.955 | 78.503 | 15.269 | 66.669 | 14.024 | 58.211 | 12.279 |
| Custom | 421.202 | 37.298 | 186.669 | 22.284 | 159.533 | 19.762 | 156.007 | 22.289 |

Table 5.2: Seed Function: Random

| Graph Modification | Avg Nodes | Nodes Std Dev | Avg Nodes C0 | Nodes Std Dev C0 | Avg Nodes C1 | Nodes Std Dev C1 | Avg Color Count | Color Count Std Dev |
|---|---|---|---|---|---|---|---|---|
| Original Graph | 103.335 | 14.119 | 14.049 | 7.982 | 14.286 | 7.729 | 11.023 | 6.184 |
| PrefAtt | 200.076 | 28.773 | 66.292 | 17.087 | 58.784 | 14.826 | 57.271 | 14.473 |
| Jaccard | 226.255 | 37.744 | 79.237 | 20.812 | 72.018 | 19.734 | 71.664 | 19.823 |
| Degree | 323.523 | 39.036 | 129.294 | 22.634 | 119.229 | 20.472 | 119.658 | 21.954 |
| TopK | 329.904 | 39.379 | 134.980 | 22.510 | 119.924 | 20.855 | 121.825 | 22.186 |
| Random | 155.483 | 18.735 | 41.669 | 10.867 | 38.814 | 10.494 | 39.695 | 9.668 |
| Custom | 368.257 | 36.911 | 158.322 | 22.137 | 134.935 | 19.110 | 143.463 | 21.419 |

Table 5.3: Seed Function: Centrality

| Graph Modification | Avg Nodes | Nodes Std Dev | Avg Nodes C0 | Nodes Std Dev C0 | Avg Nodes C1 | Nodes Std Dev C1 | Avg Color Count | Color Count Std Dev |
|---|---|---|---|---|---|---|---|---|
| Original Graph | 146.346 | 20.954 | 36.403 | 12.142 | 34.943 | 11.270 | 29.821 | 9.857 |
| PrefAtt | 232.075 | 28.991 | 86.953 | 17.397 | 70.122 | 14.897 | 74.186 | 15.209 |
| Jaccard | 235.601 | 36.974 | 84.034 | 21.053 | 76.567 | 18.865 | 75.952 | 19.807 |
| Degree | 339.339 | 37.693 | 139.271 | 21.936 | 125.068 | 19.877 | 127.002 | 21.156 |
| TopK | 340.877 | 38.641 | 141.285 | 22.171 | 124.592 | 20.447 | 127.910 | 21.915 |
| Random | 197.268 | 22.519 | 65.030 | 13.690 | 57.238 | 12.287 | 59.031 | 11.357 |
| Custom | 367.419 | 34.972 | 158.966 | 21.162 | 133.453 | 18.717 | 142.880 | 21.628 |

Table 5.4: Seed Function: Centrality Mixed

| Graph Modification | Avg Nodes | Nodes Std Dev | Avg Nodes C0 | Nodes Std Dev C0 | Avg Nodes C1 | Nodes Std Dev C1 | Avg Color Count | Color Count Std Dev |
|---|---|---|---|---|---|---|---|---|
| Original Graph | 396.762 | 22.751 | 172.331 | 15.986 | 149.431 | 13.875 | 148.103 | 13.634 |
| PrefAtt | 420.427 | 23.313 | 185.440 | 15.445 | 159.987 | 14.006 | 165.837 | 14.933 |
| Jaccard | 395.062 | 23.535 | 171.503 | 15.955 | 148.559 | 14.205 | 148.803 | 14.369 |
| Degree | 423.922 | 23.428 | 185.438 | 15.561 | 163.484 | 14.662 | 164.384 | 15.119 |
| TopK | 426.546 | 24.769 | 188.441 | 16.623 | 163.105 | 14.922 | 172.581 | 14.874 |
| Random | 426.183 | 23.832 | 187.712 | 16.282 | 163.471 | 14.085 | 169.413 | 14.901 |
| Custom | 448.260 | 23.386 | 198.499 | 15.477 | 174.761 | 14.640 | 181.655 | 15.300 |

Table 5.5: Seed Function: Degree

| Graph Modification | Avg Nodes | Nodes Std Dev | Avg Nodes C0 | Nodes Std Dev C0 | Avg Nodes C1 | Nodes Std Dev C1 | Avg Color Count | Color Count Std Dev |
|---|---|---|---|---|---|---|---|---|
| Original Graph | 148.891 | 21.033 | 46.603 | 12.231 | 27.288 | 11.144 | 23.356 | 8.854 |
| PrefAtt | 227.877 | 28.664 | 84.514 | 16.384 | 68.363 | 15.470 | 60.225 | 13.291 |
| Jaccard | 322.371 | 41.937 | 137.751 | 23.452 | 109.620 | 21.988 | 103.214 | 21.358 |
| Degree | 340.958 | 39.268 | 142.552 | 21.780 | 123.406 | 21.523 | 113.846 | 21.411 |
| TopK | 385.825 | 40.055 | 164.234 | 21.909 | 146.591 | 22.318 | 134.513 | 22.118 |
| Random | 218.832 | 25.798 | 81.064 | 14.953 | 62.768 | 13.869 | 57.950 | 12.002 |
| Custom | 421.036 | 39.374 | 182.475 | 21.803 | 163.561 | 21.847 | 152.176 | 22.983 |

Table 5.6: Seed Function: Polarized

| Graph Modification | Avg Nodes | Nodes Std Dev | Avg Nodes C0 | Nodes Std Dev C0 | Avg Nodes C1 | Nodes Std Dev C1 | Avg Color Count | Color Count Std Dev |
|---|---|---|---|---|---|---|---|---|
| Original Graph | 100.017 | 13.407 | 13.590 | 7.366 | 11.427 | 7.474 | 9.347 | 5.807 |
| PrefAtt | 193.429 | 27.604 | 61.201 | 14.472 | 57.228 | 15.847 | 48.728 | 13.050 |
| Jaccard | 235.383 | 40.186 | 78.822 | 20.515 | 81.561 | 22.202 | 74.501 | 20.847 |
| Degree | 311.227 | 38.870 | 111.559 | 20.693 | 124.668 | 21.589 | 111.898 | 22.086 |
| TopK | 336.932 | 40.287 | 125.984 | 21.319 | 135.948 | 22.768 | 121.664 | 22.859 |
| Random | 162.856 | 19.452 | 48.280 | 10.612 | 39.576 | 11.408 | 35.304 | 9.140 |
| Custom | 374.008 | 38.502 | 146.914 | 20.964 | 152.094 | 21.637 | 138.47 | 22.748 |

Table 5.7: Seed Function: Polarized Centrality

| Graph Modification | Avg Nodes | Nodes Std Dev | Avg Nodes C0 | Nodes Std Dev C0 | Avg Nodes C1 | Nodes Std Dev C1 | Avg Color Count | Color Count Std Dev |
|---|---|---|---|---|---|---|---|---|
| Original Graph | 179.565 | 23.584 | 52.158 | 12.821 | 52.407 | 13.848 | 42.731 | 11.011 |
| PrefAtt | 276.223 | 33.380 | 95.708 | 17.779 | 105.515 | 19.172 | 93.457 | 17.844 |
| Jaccard | 256.401 | 37.013 | 88.222 | 18.819 | 93.179 | 21.057 | 84.341 | 19.569 |
| Degree | 344.449 | 36.982 | 127.105 | 19.411 | 142.344 | 21.546 | 129.024 | 21.625 |
| TopK | 360.382 | 36.762 | 138.246 | 20.116 | 147.136 | 21.184 | 133.59 | 21.723 |
| Random | 254.827 | 29.532 | 92.095 | 15.565 | 87.732 | 17.516 | 78.898 | 15.559 |
| Custom | 392.172 | 34.577 | 154.430 | 19.109 | 162.742 | 19.942 | 150.19 | 21.347 |

Table 5.8: Seed Function: Polarized Centrality Mixed

| Graph Modification | Avg Nodes | Nodes Std Dev | Avg Nodes C0 | Nodes Std Dev C0 | Avg Nodes C1 | Nodes Std Dev C1 | Avg Color Count | Color Count Std Dev |
|---|---|---|---|---|---|---|---|---|
| Original Graph | 304.438 | 24.934 | 122.591 | 13.899 | 106.847 | 15.747 | 95.464 | 14.030 |
| PrefAtt | 365.123 | 28.988 | 152.709 | 16.507 | 137.414 | 17.572 | 128.137 | 17.309 |
| Jaccard | 350.439 | 31.186 | 144.025 | 17.278 | 131.414 | 18.552 | 124.546 | 18.816 |
| Degree | 396.485 | 30.408 | 157.023 | 16.291 | 164.462 | 19.268 | 153.975 | 19.651 |
| TopK | 398.274 | 30.998 | 159.496 | 17.055 | 163.778 | 19.190 | 152.923 | 20.080 |
| Random | 344.710 | 25.241 | 140.715 | 14.454 | 128.995 | 15.963 | 120.812 | 15.220 |
| Custom | 429.097 | 31.319 | 173.603 | 17.390 | 180.494 | 19.249 | 172.163 | 20.596 |

Table 5.9: Seed Function: Polarized Degree

## 5.6 Analysis of Tabular Results

The following tables present comprehensive results from experiments conducted on a synthetic NETIN network consisting of 1500 nodes and 89.9k edges, with a homophily value of $h = 0.8$. These tables provide detailed metrics across different seed functions and modification strategies, including average node activation, color-specific performance (C0 and C1), and their respective standard deviations. The high homophily value ($h = 0.8$) indicates a strong community structure, making it particularly suitable for evaluating the effectiveness of various graph modification strategies in overcoming community boundaries.

### 5.6.1 Performance Across Seed Functions

The tabular results provide comprehensive insights into the performance of different graph modification strategies across various seed functions. For the Random seed function (Table 5.2), the Custom modification strategy demonstrates superior performance with an average node activation of 421.202 nodes (±37.298), significantly outperforming other approaches. The TopK strategy follows as the second-best performer with 385.691 nodes (±39.598), while the Random modification shows the lowest performance among the modified graphs with only 220.172 nodes (±25.955).

When examining the Centrality-based seed function (Table 5.3), we observe a different pattern of performance. While Custom maintains its leadership with 368.257 nodes (±36.911), the gap between Custom and TopK (329.904 ±39.379) narrows compared to the Random seed function. This indicates that the effectiveness of modification strategies is influenced by the initial seed selection method.

### 5.6.2 Color Distribution Analysis

The color activation metrics reveal interesting patterns across different configurations. In the Polarized setup (Table 5.6), Custom achieves the most balanced color distribution between C0 (182.475 ±21.803) and C1 (163.561 ±21.847), with an impressive average color count of 152.176 (±22.983). This balanced performance is particularly noteworthy given the inherent challenges of polarized configurations.

The Polarized Centrality configuration (Table 5.7) shows even more striking results in terms of color balance. Custom maintains strong performance with 374.008 total nodes (±38.502) while achieving near-perfect balance between C0 (146.914 ±20.964) and C1 (152.094 ±21.637). This indicates the strategy's robustness in handling polarized network structures while maintaining effective cross-community activation.

### 5.6.3 Comparative Strategy Analysis

A cross-comparison of strategies reveals several key insights:

1. **Custom Strategy Superiority:** Across all seed functions, the Custom modification strategy consistently achieves the highest average node activation and color count metrics. This superiority is particularly evident in Table 5.5, where Custom reaches

448.260 nodes (±23.386) compared to the next best performer, TopK, at 426.546 nodes (±24.769).

2. **Baseline Performance:** The Original Graph configuration serves as a baseline, consistently showing the lowest performance across all metrics. This is most evident in Table 5.3, where the original graph achieves only 103.335 nodes (±14.119) compared to Custom's 368.257 nodes (±36.911).

3. **Strategy Consistency:** The relative performance ranking of strategies remains fairly consistent across different seed functions, with Custom > TopK > Degree > Jaccard > PrefAtt > Random being the typical order. This pattern is particularly clear in Table 5.2 and Table 5.4.

### 5.6.4 Standard Deviation Analysis

The standard deviation patterns provide insights into strategy stability:

1. Modified graphs generally show higher standard deviations compared to the original graph, indicating increased variability in outcomes. This is particularly noticeable in Table 5.8, where Custom shows a standard deviation of 34.577 compared to the original graph's 23.584.

2. The Custom strategy, despite its higher average performance, maintains comparable or sometimes lower standard deviations relative to other modification strategies, suggesting reliable performance improvements.

### 5.6.5 Mixed Configuration Performance

The mixed configuration results (Table 5.4 and Table 5.8) demonstrate the strategies' adaptability to hybrid scenarios. Custom maintains its performance advantage while showing robust handling of the increased complexity introduced by mixed configurations. This is evidenced by the high average color count (142.880 ±21.628) in Table 5.4, suggesting effective cross-community activation even in these more challenging scenarios.

These comprehensive results validate the superior performance of the Custom modification strategy across various network configurations and seed functions while also highlighting the importance of considering both node activation and color distribution metrics in evaluating strategy effectiveness.

# 6 Conclusion

In this thesis, we investigated influence maximization techniques aimed at minimizing polarization in social networks through strategic graph modification strategies. Through extensive experimentation on both real-world networks (Twitter Interaction Network for the US Congress and AVES-WEAVER-SOCIAL) and synthetic networks of varying sizes and homophily values, we demonstrated that our Custom modification strategy consistently outperforms traditional approaches in both influence spread and polarization reduction.

The empirical results across different network configurations reveal that Custom modifications achieve superior activation rates, particularly evident in the Twitter network with 475 nodes and 13.2k edges, where they consistently outperform baseline methods by $20-30\%$. This performance advantage persists across different seed functions, with notably strong results in Degree and Centrality-based seed selections. The strategy's effectiveness is particularly pronounced in networks with high homophily ($h = 0.8$), where it successfully bridges strongly defined community structures.

A key finding emerges from the comparison of different network sizes and structures: while Custom maintains its performance advantage across all tested configurations, the magnitude of improvement varies with network characteristics. In the smaller AVES network (445 nodes, 1.4k edges), we observe steeper initial activation curves but earlier plateaus, while larger synthetic networks (1000 nodes) demonstrate more gradual but sustained improvement at higher k-values. This scalability is particularly noteworthy as it suggests the strategy's potential applicability to larger real-world networks.

### 6.0.1 Limitations of This Work

While our study provides substantial evidence for the effectiveness of Custom modifications, several limitations should be acknowledged:

- The evaluation, while comprehensive across different network sizes, remains bounded by computational constraints in testing extremely large networks
- The Independent Cascade Model serves as our primary diffusion mechanism, leaving room for exploration of alternative models
- While we tested two real-world networks, additional validation across diverse network types would strengthen generalizability
- The computational complexity of Custom modifications may pose challenges for real-time applications in rapidly evolving networks

### 6.0.2 Future Work

Several promising directions for future research emerge from our findings:

- Investigation of dynamic network modifications, particularly relevant for real-world social networks where relationships evolve over time
- Development of optimized implementations to reduce computational overhead while maintaining modification effectiveness
- Extension to alternative diffusion models, including threshold-based and hybrid approaches
- Exploration of adaptive strategies that automatically adjust to different network structures and community configurations
- Analysis of the long-term stability of modified network structures and their impact on sustained influence spread
- Investigation of the strategy's effectiveness in networks with more complex community structures and multiple competing influences

The strong performance of Custom modifications across diverse network configurations suggests significant potential for practical applications in social network analysis and influence maximization. While computational efficiency remains a challenge, particularly for larger networks, the consistent ability to enhance influence spread while bridging community divisions marks an important advancement in addressing network polarization. Future work focusing on optimization and scalability could further enhance the practical applicability of these strategies in real-world scenarios.

# Bibliography


[1] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[2] Çigdem Aslay, Antonis Matakos, Esther Galbrun, and Aristides Gionis. Maximizing the diversity of exposure in a social network. *CoRR*, abs/1809.04393, 2018.

[3] Lars Backstrom and Jure Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, page 635–644, New York, NY, USA, 2011. Association for Computing Machinery.

[4] Ruben Becker, Federico Corò, Gianlorenzo D'Angelo, and Hugo Gilbert. Balancing spreads of influence in a social network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):3–10, Apr. 2020.

[5] Elisabetta Bergamini, Pierluigi Crescenzi, Gianlorenzo D'angelo, Henning Meyerhenke, Lorenzo Severini, and Yllka Velaj. Improving the betweenness centrality of a node by adding links. *ACM J. Exp. Algorithmics*, 23, aug 2018.

[6] Shishir Bharathi, David Kempe, and Mahyar Salek. Competitive influence maximization in social networks. In Xiaotie Deng and Fan Chung Graham, editors, *Internet and Network Economics*, pages 306–311, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[7] Paolo Boldi and Sebastiano Vigna. Axioms for centrality. *CoRR*, abs/1308.2140, 2013.

[8] Vineet Chaoji, Sayan Ranu, Rajeev Rastogi, and Rushi Bhatt. Recommendations to boost content spread in social networks. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 529–538, New York, NY, USA, 2012. Association for Computing Machinery.

[9] Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, page 1029–1038, New York, NY, USA, 2010. Association for Computing Machinery.

[10] Wei Chen, Yifei Yuan, and Li Zhang. Scalable influence maximization in social networks under the linear threshold model. In *2010 IEEE International Conference on Data Mining*, pages 88–97, 2010.

[11] Edith Cohen, Daniel Delling, Thomas Pajor, and Renato F. Werneck. Sketch-based influence maximization and computation: Scaling up with guarantees. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, page 629–638, New York, NY, USA, 2014. Association for Computing Machinery.

[12] Federico Coró, Gianlorenzo D'angelo, and Yllka Velaj. Link recommendation for social influence maximization. *ACM Trans. Knowl. Discov. Data*, 15(6), may 2021.

[13] Alessandro Cossard, Gianmarco De Francisci Morales, Kyriaki Kalimeri, Yelena Mejova, Daniela Paolotti, and Michele Starnini. Falling into the echo chamber: the italian vaccination debate on twitter. *CoRR*, abs/2003.11906, 2020.

[14] Pierluigi Crescenzi, Gianlorenzo D'angelo, Lorenzo Severini, and Yllka Velaj. Greedily improving our own closeness centrality in a network. *ACM Trans. Knowl. Discov. Data*, 11(1), jul 2016.

[15] Gianlorenzo D'Angelo, Lorenzo Severini, and Yllka Velaj. Recommending links through influence maximization. *Theoretical Computer Science*, 764:30–41, 2019. Selected papers of ICTCS 2016 (The Italian Conference on Theoretical Computer Science (ICTCS).

[16] Maria Deijfen and Mathias Lindholm. Growing networks with preferential deletion and addition of edges. *Physica A: Statistical Mechanics and its Applications*, 388(19):4297–4303, October 2009.

[17] Erik D. Demaine and Morteza Zadimoghaddam. Minimizing the diameter of a network using shortcut edges. In Haim Kaplan, editor, *Algorithm Theory - SWAT 2010*, pages 420–431, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[18] Magdalini Eirinaki, Jerry Gao, Iraklis Varlamis, and Konstantinos Tserpes. Recommender systems for large-scale social networks: A review of challenges and solutions. *Future Generation Computer Systems*, 78:413–418, 2018.

[19] Lisette Espin-Noboa Fariba Karimi and Jan Bachmann. netin: Network inequality. `https://cshvienna.github.io/NetworkInequalities/alpha/`, 2024. Version 2.0.0a1.

[20] Christian G Fink, Kelly Fullin, Guillermo Gutierrez, Nathan Omodt, Sydney Zinnecker, Gina Sprint, and Sean McCulloch. A centrality measure for quantifying spread on weighted, directed networks. *Physica A*, 2023.

[21] Seth Flaxman, Sharad Goel, and Justin M. Rao. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, 80(S1):298–320, 03 2016.

[22] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[23] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Reducing controversy by connecting opposing views, 2018.

[24] Shahrzad Haddadan, Cristina Menghini, Matteo Riondato, and Eli Upfal. Reducing polarization and increasing diverse navigability in graphs by inserting edges and swapping edge weights. *Data Mining and Knowledge Discovery*, 36(6):2334–2378, 2022.

[25] Ruben Interian, Jorge R. Moreno, and Celso C. Ribeiro. Polarization reduction by minimum-cardinality edge additions: Complexity and integer programming approaches. *International Transactions in Operational Research*, 28(3):1242–1264, 2021.

[26] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, page 137–146, New York, NY, USA, 2003. Association for Computing Machinery.

[27] Antonis Matakos, Sijing Tu, and Aristides Gionis. Tell me something my friends do not know: diversity maximization in social networks. *Knowledge and Information Systems*, 62(9):3697–3726, 2020.

[28] Cameron Musco, Christopher Musco, and Charalampos E. Tsourakakis. Minimizing polarization and disagreement in social networks, 2017.

[29] Nikos Parotsidis, Evaggelia Pitoura, and Panayiotis Tsaparas. *Selecting Shortcuts for a Smaller World*, pages 28–36.

[30] Ryan A. Rossi and Nesreen K. Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI*, 2015.

[31] David Shultz. Could google influence the presidential election? *Science*, October 2016. Accessed: 2024-05-11.

[32] Liwang Zhu, Qi Bao, and Zhongzhi Zhang. Minimizing polarization and disagreement in social networks via link recommendation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2072–2084. Curran Associates, Inc., 2021.

[33] Antonio Zinilli and Giovanni Cerulli. Link prediction and feature relevance in knowledge networks: A machine learning approach. *PLOS ONE*, 18(11):1–32, 11 2023.