**Dynamic Analysis of Information Diffusion in Social Networks: Evidence from Facebook**

Yiwei Qi, Hongzhou Tao

Graduate School of Arts and Sciences, Georgetown University, USA

10个研究问题：  
 **社交网络的度分布特征是什么？**

* **背景**：了解度分布可以揭示节点在网络中的连接数，帮助识别网络中是否存在超级节点。
* **数据文件**：.edges 文件
* **方法**：构建社交网络图，计算度分布。

 **社交网络的聚类系数和平均路径长度是什么？**

* **背景**：聚类系数和平均路径长度可以描述网络的整体结构特征和紧密度。
* **数据文件**：.edges 文件
* **方法**：构建社交网络图，计算聚类系数和平均路径长度。

 **在社交网络中是否存在超级节点（超级传播者）？**

* **背景**：识别超级节点可以帮助理解哪些节点在信息扩散中起关键作用。
* **数据文件**：.edges 文件
* **方法**：通过度分布和中心性指标识别超级节点。

 **如何使用社交网络中的节点和边数据建模信息扩散过程？**

* **背景**：建立信息扩散模型可以模拟信息在网络中的传播路径和模式。
* **数据文件**：.edges 文件，结合用户属性数据
* **方法**：构建信息扩散模型，模拟信息传播。

 **不同类型的信息（如新闻、谣言、广告）在社交网络中的扩散模式有何不同？**

* **背景**：了解不同类型信息的扩散模式有助于制定有针对性的传播策略。
* **数据文件**：.edges 文件，结合用户属性数据
* **方法**：模拟不同类型信息的扩散过程，比较扩散模式。

 **如何衡量用户在社交网络中的影响力？**

* **背景**：通过度中心性、介数中心性和接近中心性等指标可以量化用户的影响力。
* **数据文件**：.edges 文件，结合用户属性数据
* **方法**：计算用户的中心性指标。

 **高影响力用户在信息扩散过程中起到了什么作用？**

* **背景**：高影响力用户可能加速或减缓信息扩散的速度和范围。
* **数据文件**：.edges 文件，结合用户属性数据
* **方法**：识别高影响力用户，分析其在信息扩散中的作用。

 **不同类型的信息在社交网络中的扩散范围和速度有何不同？**

* **背景**：比较不同类型信息的扩散特征可以揭示其传播效率和覆盖范围。
* **数据文件**：Graph Embedding with Self Clustering
* **方法**：模拟信息扩散过程，记录扩散范围和速度。

 **如何优化社交网络中的信息传播路径，以最小化传播时间或最大化传播范围？**

* **背景**：优化信息传播路径可以提高信息传递的效率。
* **数据文件**：Facebook Large Page-Page Network
* **方法**：使用网络优化算法优化传播路径，比较不同策略的效果。

 **优化前后的信息传播效率有何差异？**

* **背景**：通过比较优化前后的传播效率，可以评估优化策略的有效性。
* **数据文件**：Facebook Large Page-Page Network
* **方法**：分析优化前后的信息传播效率，使用可视化方法展示优化后的传播路径。

ABSTRACT – 内容、目的、方法、发现和结论

Keywords

1. Introduction

1. 研究背景
2. 研究问题与目标
3. 本文结构

2.Literature Review

要求紧紧围绕研究问题与目标总结已有研究

3.实验设计

数据收集、数据集描述与预处理

模型构建：描述不同结构模型的基本原理与方法

实验使用的方法与参数设置等

4.实验结果与讨论

4.1对网络模型的分析

1. 社交网络结构分析

- \*\*研究问题\*\*：整个社交网络的度分布、聚类系数和平均路径长度等网络结构特征是什么？是否存在超级节点（超级传播者）？

- \*\*数据文件\*\*：.edges 文件

- \*\*逻辑顺序\*\*：

- 构建社交网络图

- 计算网络结构指标

- 识别超级节点

2. 信息扩散模型

- \*\*研究问题\*\*：如何使用社交网络中的节点和边数据建模信息扩散过程？不同类型的信息（如新闻、谣言、广告）在社交网络中的扩散模式有何不同？

- \*\*数据文件\*\*：.edges 文件，结合用户属性数据

- \*\*逻辑顺序\*\*：

- 构建信息扩散模型

- 模拟信息在网络中的传播

- 比较不同类型信息的扩散模式

3. 用户影响力分析

- \*\*研究问题\*\*：如何衡量用户在社交网络中的影响力（如通过度中心性、介数中心性和接近中心性等指标）？高影响力用户在信息扩散过程中起到了什么作用？

- \*\*数据文件\*\*：.edges 文件，结合用户属性数据

- \*\*逻辑顺序\*\*：

- 计算用户的中心性指标

- 识别高影响力用户

- 分析高影响力用户在信息扩散中的作用

4. **研究问题 4: 不同类型信息的扩散模式分析**

**课程相关性**: 模块4中的“Network Dynamics”章节

**研究问题**:

* 不同类型的信息（如新闻、谣言、广告）在社交网络中的扩散模式有何不同？

**使用的数据集**:

* Graph Embedding with Self Clustering

**逻辑顺序**:

1. **数据分类**:
   * 根据页面的描述信息，分类不同类型的信息（如新闻、谣言、广告）。
2. **扩散过程模拟**:
   * 在网络中模拟不同类型信息的扩散过程。
   * 记录不同类型信息的扩散范围和速度。
3. **模式对比**:
   * 比较不同类型信息的扩散模式，确定其异同点。
   * 使用可视化方法展示不同类型信息的扩散路径。

5. **社交网络中的优化问题**

**课程相关性**: 模块5中的“Network Optimization”章节

**研究问题**:

* 如何优化社交网络中的信息传播路径，以最小化传播时间或最大化传播范围？

**使用的数据集**:

* Facebook Large Page-Page Network

**逻辑顺序**:

1. **数据预处理**:
   * 提取节点和边的信息，标记节点的类别。
2. **路径优化**:
   * 使用网络优化算法（如最短路径算法、最大流算法）优化信息传播路径。
   * 比较不同优化策略的效果。
3. **优化分析**:
   * 分析优化前后的信息传播效率。
   * 使用可视化方法展示优化后的信息传播路径。

5. 结论

再复述下结论：用一大段话回答10个问题 注意内容包括：

动态机制分析

总结模拟实验和数据分析的结果，揭示信息扩散的动态机制。

讨论影响扩散结果的关键因素。

展望未来发展方向

参考文献

Abstract

This study delves into the dynamics of information dissemination in social networks, focusing on the impact of different network structures. We analyze various network types to understand their influence on the spread speed and pattern of information. Our research includes using Facebook, constructing social network graphs, calculating network structure metrics, and identifying supernodes. We also model the information diffusion process, comparing the diffusion patterns of different types of information such as news, rumors, and advertisements. Our analysis reveals that network structure significantly influences the spread of information. High connectivity accelerates dissemination, while excessive clustering may decelerate it. Our findings highlight the importance of considering network structure when formulating information dissemination strategies.

Keywords: Information Diffusion, Social Networks, Network Analysis, Information Propagation

1. INTRODUCTION

In today’ s digital era, the spread of information in social networks has become a key research topic, profoundly impacting fields such as marketing, public health, and politics. The rapid proliferation of digital platforms and social media has made it easier for information to spread quickly and widely, reaching a large audience in a short amount of time. Understanding the diffusion mechanism of information in the network not only helps optimize communication strategies but also predicts the coverage and influence of different types of content. This understanding is crucial for improving the efficiency and accuracy of information exchange, ensuring that the right information reaches the right people at the right time.

The phenomenon of information dissemination in social networks is complex and diverse. Different network structures, such as random networks, small-world networks, and scale-free networks, have a significant impact on information diffusion. Each of these structures has unique characteristics that influence the way and speed of information propagation. For example, random networks, characterized by a uniform probability of connections between nodes, may exhibit different diffusion patterns compared to small-world networks, which feature high clustering and short average path lengths. Scale-free networks, with their few highly connected nodes (hubs), present yet another set of diffusion dynamics. Therefore, studying information dissemination in social networks can help reveal the dynamic mechanisms and key influencing factors behind it, providing a deeper understanding of how information flows through these complex systems.

The significance of studying information dissemination lies in its wide range of practical applications. Research in this area can help design effective marketing strategies, manage the spread of disinformation, and respond quickly in public health emergencies. For example, in times of crisis, understanding the dissemination patterns of information can ensure that critical information is conveyed to the public quickly and effectively, thus improving the efficiency and effectiveness of emergency response. In addition, accurately understanding the laws of information dissemination can help various organizations formulate more targeted communication strategies and improve the arrival rate and influence of information.

In marketing, companies can develop more attractive and influential advertising strategies to increase the exposure and awareness of products or services by understanding the mechanism of information diffusion. This can lead to better-targeted marketing campaigns, higher conversion rates, and improved customer engagement. In public health, mastering information dissemination patterns helps the government and relevant agencies release accurate information in a timely manner, reducing the impact of rumors and misinformation in emergencies such as disease outbreaks, thereby protecting public health. Effective communication strategies can ensure that health advisories and preventive measures reach the population swiftly, thereby mitigating the spread of diseases. In politics, understanding the dynamics of information dissemination can help candidates and parties develop effective campaign strategies to win over more voters. By analyzing how political messages spread through social networks, campaigns can tailor their strategies to maximize reach and influence, ensuring that their messages resonate with the target audience.

Furthermore, the study of information dissemination is critical for combating the spread of misinformation and fake news. In an age where false information can spread rapidly through social networks, understanding how to control and counteract such dissemination is essential for maintaining the integrity of information. This research can inform the development of algorithms and policies that identify and mitigate the spread of false information, ensuring that accurate and reliable information prevails.

The main goal of this study is to analyze the dynamic characteristics of information diffusion in social networks and explore how different network structures affect the speed and pattern of information diffusion. By employing advanced modeling techniques, we aim to reveal the underlying mechanisms of information dissemination and identify the key factors that influence the dissemination results. Specifically, we will conduct simulation experiments and data analysis to study the characteristics of information dissemination in different network environments. These simulations will help us understand how various network topologies and user interactions contribute to the spread of information.

The structure of this paper is designed to systematically explore the issue of information diffusion in social networks. Chapter 2, Literature Review, summarizes existing research to provide a theoretical foundation. Chapter 3, Experimental Design, describes data collection, preprocessing, model construction, and experimental methods and parameters. Chapter 4, Experimental Results and Discussion, covers social network structure analysis, information diffusion modeling, user influence analysis, the study of different information diffusion patterns, and methods for optimizing information dissemination paths. Chapter 5, Conclusion, summarizes the results, reveals the dynamic mechanisms of information diffusion, discusses key influencing factors, and looks ahead to future research directions.

1. LITERATURE REVIEW

Newman (2003) in "Structure and Function of Complex Networks" highlighted the significant impact of network topology on information dissemination, emphasizing the important role of small-world effect and scale-free characteristics in promoting rapid dissemination. By analyzing nodes and edges in the network structure, he revealed how high clustering and short path lengths in small-world networks promote rapid propagation of information in local clusters, and discussed the role of a few highly connected nodes in scale-free networks in the widespread but uneven diffusion of information. These studies provide a foundation for understanding degree distribution characteristics and identifying super nodes. Beyond degree distribution, clustering coefficient and average path length of the network also significantly impact information diffusion. Xu and Shen (2015) analyzed the significant impact of opportunistic links and node mobility on information dissemination in mobile social networks, revealing through simulation experiments how these factors change the path and speed of information dissemination, thus affecting the overall communication effect. These studies offer perspectives for understanding the overall structural characteristics and tightness of the network.

The construction and simulation of information diffusion models are crucial in the study of information dissemination. Morales and Benito (2014) studied the efficiency of information dissemination on Twitter and found that active and influential users play a key role in the diffusion process. By analyzing Twitter data, they pointed out that users with a large number of followers and who frequently post information can significantly accelerate information dissemination, which is significant for designing targeted marketing and communication strategies. Liu et al. (2019), in "Nonlinear Dynamic Modeling and Evolutionary Analysis of Information Propagation in Online Social Networks," discussed the nonlinear dynamics of information propagation under emergencies, constructing a nonlinear dynamic model to analyze the information propagation mode under emergencies and revealing the key influencing factors in the propagation process. These studies provide a theoretical foundation for constructing information diffusion models and simulating information dissemination.

Different types of information may exhibit significant differences in their diffusion patterns in social networks. Zhang and Zhang (2016) discussed the practical application of diffusion models in fields such as disease spread, network security, and public opinion monitoring, pointing out that diffusion models are of great significance in both theory and practical applications. By applying them to different fields, information dissemination patterns can be better understood and predicted, and more effective response strategies can be developed. These studies provide a basis for understanding the diffusion characteristics of different types of information and formulating dissemination strategies.

User influence analysis is also indispensable in the study of information dissemination. Jiang and Liu (2014) used an evolutionary game theory-based model to reveal the dynamic changes in communication strategies in social networks. They constructed an evolutionary game model to analyze the evolution process of different communication strategies in the network, revealing users' strategy choices in information communication and their impact on diffusion results, which is significant for understanding the complex dynamics of information dissemination in social networks. Chatman (1986) emphasized that information dissemination depends on network structure, information content, and the cognitive state of the recipient. He pointed out that network structure determines the path and speed of information dissemination, information content affects dissemination effectiveness, and the cognitive state of the receiver determines the acceptance and willingness to disseminate information, providing theoretical support for understanding multiple factors in information dissemination. These studies offer methods and indicators for measuring user influence in social networks.

In summary, existing research indicates that network structure and user behavior play crucial roles in information dissemination, but many issues remain to be explored. Based on previous work, this study will further explore the impact of different network structures and user behaviors on information dissemination, aiming to reveal the dynamic mechanisms of information dissemination and provide theoretical support and practical guidance for optimizing dissemination strategies.

1. EXPERIMENTAL DESIGN

This study uses three main datasets: Social Circles: Facebook, Graph Embedding with Self Clustering, and Facebook Large Page-Page Network from SNAP (https://snap.stanford.edu/index.html). These datasets provide rich node and edge data, enabling us to deeply analyze the characteristics of information diffusion and optimization strategies in social networks. Data statistics are shown as Table 1 below.

Table 1: Data Statistics

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Social circles: Facebook | | Graph Embedding with Self Clustering: Facebook | | |  | | Facebook Large Page-Page Network | |
| Category | | Nodes | | Edges |
| Nodes in largest WCC | 4039 (1.000) | Government | 7,057 | | 89,455 | | Directed | No. |
| Edges in largest WCC | 88234 (1.000) | New Sites | 27,917 | | 206,259 | | Node features | Yes. |
| Nodes in largest SCC | 4039 (1.000) | Athletes | 13,866 | | 86,858 | | Edge features | No. |
| Edges in largest SCC | 88234 (1.000) | Public Figures | 11,565 | | 67,114 | | Node labels | Yes. Binary-labeled. |
| Average clustering coefficient | 0.6055 | TV Shows | 3,892 | | 17,262 | | Temporal | No. |
| Nodes | 4039 | Politician | 5,908 | | 41,729 | | Nodes | 22,470 |
| Edges | 88234 | Artist | 50,515 | | 819,306 | | Edges | 171,002 |
| Number of triangles | 1612010 | Company | 14,113 | | 52,310 | | Density | 0.001 |
| Fraction of closed triangles | 0.2647 |  |  | |  | | Transitvity | 0.232 |
| Diameter (longest shortest path) | 8 |  |  | |  | |  |  |
| 90-percentile effective diameter | 4.7 |  |  | |  | |  |  |

The experiment first analyzes the overall characteristics of the network by constructing a social network graph, calculating network structure indicators such as degree distribution, clustering coefficient and average path length, and identifying super nodes. The specific method includes using .edges files to construct a social network graph, calculating the degree distribution, clustering coefficient and average path length of the network. When identifying super nodes, degree distribution and centrality indicators (such as degree centrality, betweenness centrality and closeness centrality) are used for analysis. To simulate the information diffusion process, we use the SIR (susceptible-infected-recovered) model, which includes three states: susceptible (S), infected (I) and recovered (R), and uses infection rate (β) and recovery rate (γ) as model parameters. By randomly selecting the initial infected node and setting appropriate parameters, we simulated the information propagation process in the social network, and simulated three types of information: news, rumors and advertisements, and compared their propagation speed and coverage.

In the user influence analysis, we measure the influence of users in the social network by calculating indicators such as degree centrality, betweenness centrality and closeness centrality. Degree centrality reflects the number of connections of a node, betweenness centrality indicates the importance of a node in the shortest path, and closeness centrality indicates the average shortest path length from a node to other nodes. Through these indicators, we identified highly influential users in the network and analyzed their role in the information diffusion process. Users with high degree centrality have a large number of direct connections and are the main nodes for information dissemination; users with high betweenness centrality play a bridging role in the path of information transmission, connecting different parts of the network; users with high closeness centrality can quickly reach other users in the network and are important nodes for rapid information dissemination.

In terms of optimizing the information dissemination path, we used the Dijkstra algorithm and the Edmonds-Karp algorithm to optimize the dissemination path. The Dijkstra algorithm is used to calculate the shortest path, while the Edmonds-Karp algorithm is used to calculate the maximum flow. By comparing the effects of different optimization strategies, we evaluated the information dissemination efficiency before and after optimization, and used visualization methods to display the optimized dissemination path. The construction of these methods and models enables us to deeply explore the characteristics of information diffusion in social networks, compare the diffusion patterns of different types of information, and optimize the information dissemination path, ultimately providing theoretical support and practical guidance for the formulation of effective dissemination strategies.

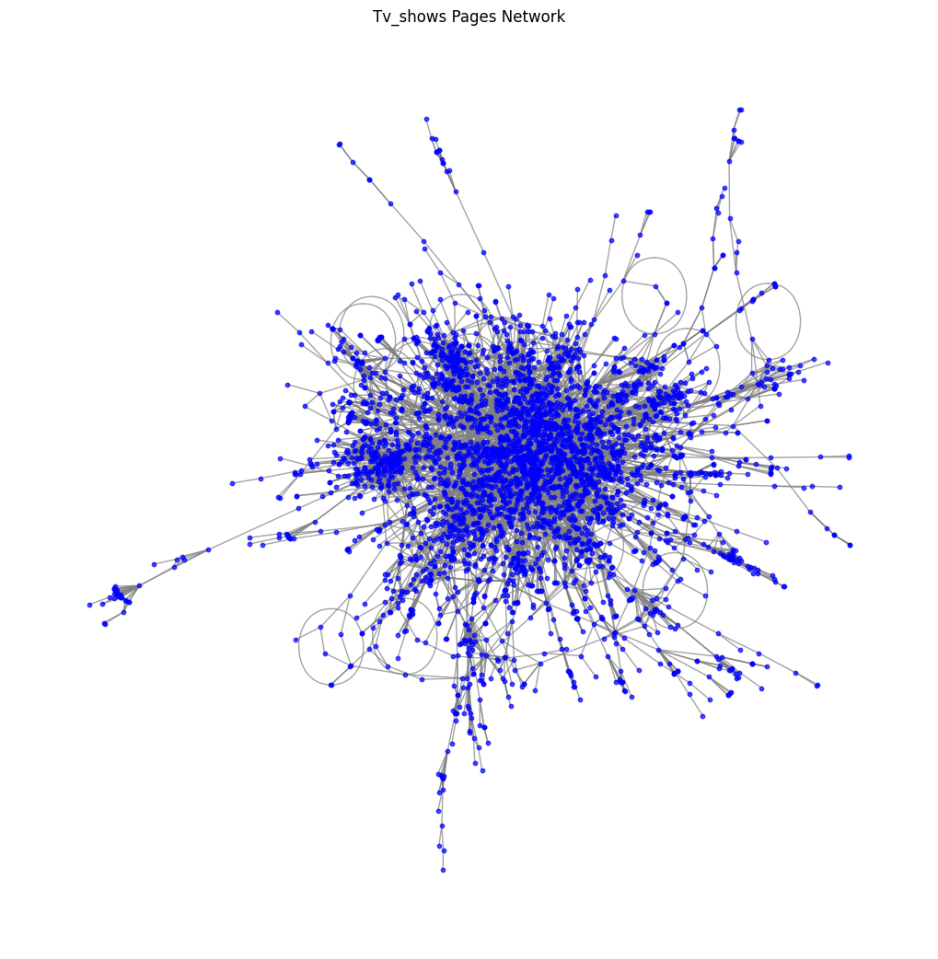
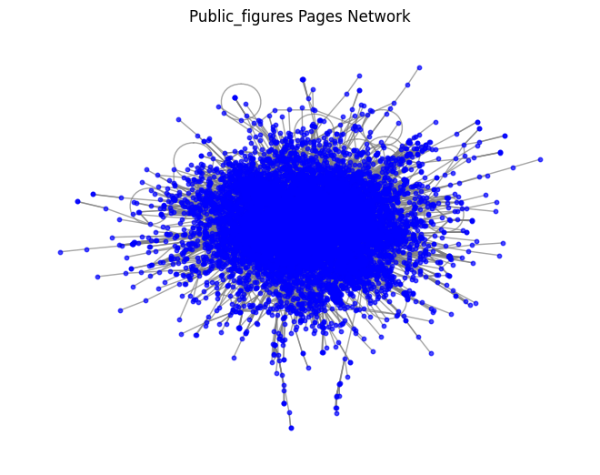
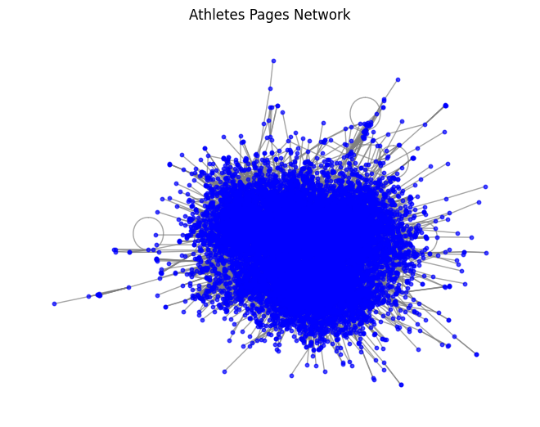
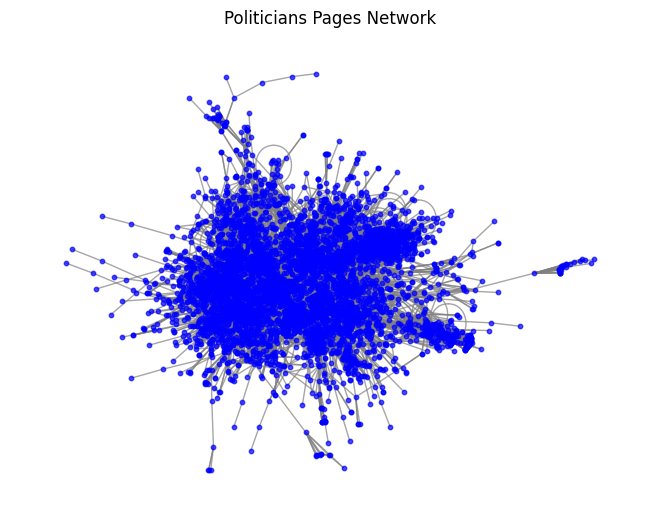
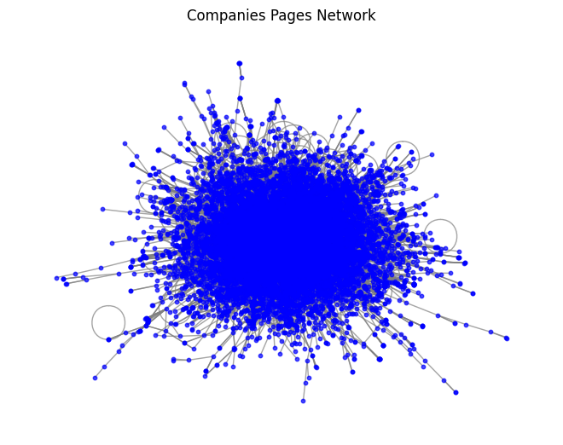
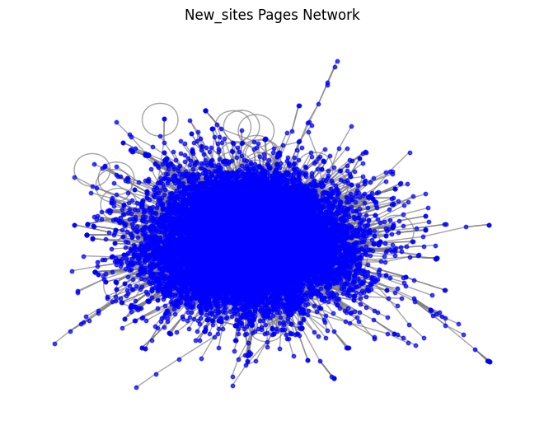
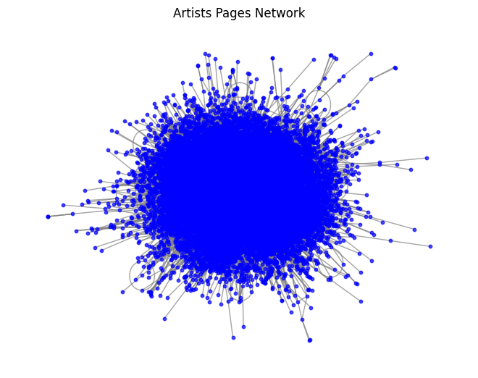
1. EXPERIMENT RESULTS

4.1

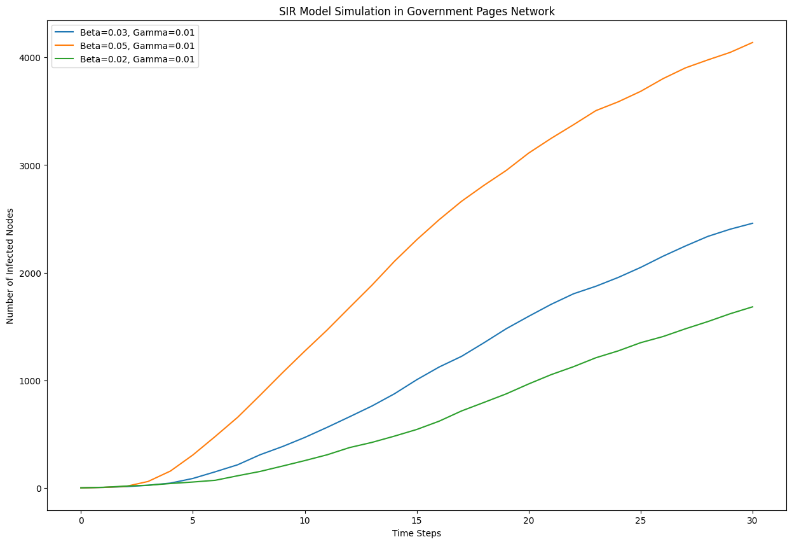
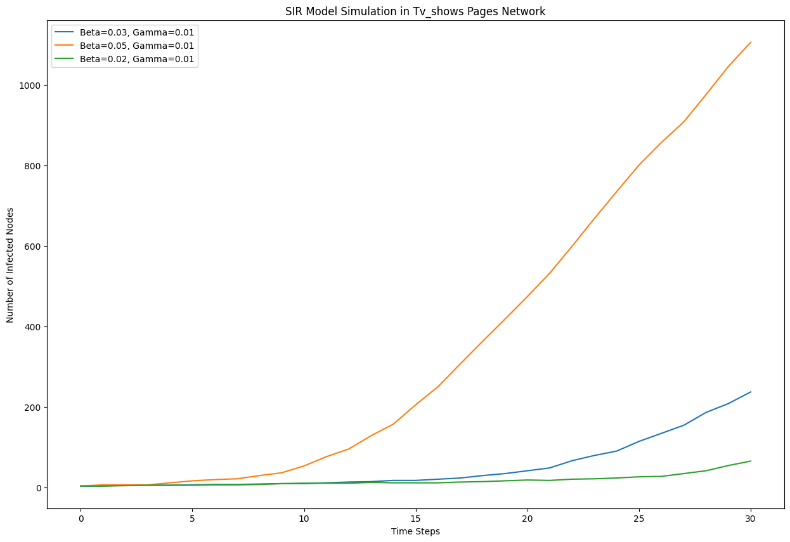
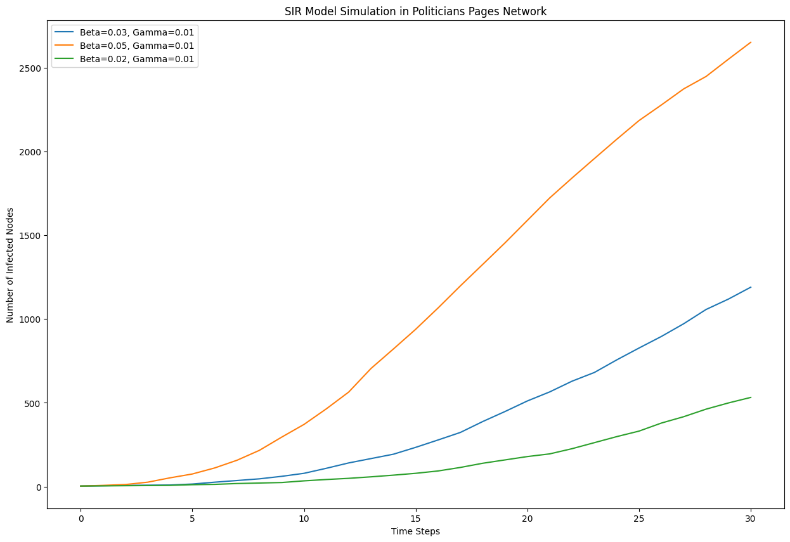
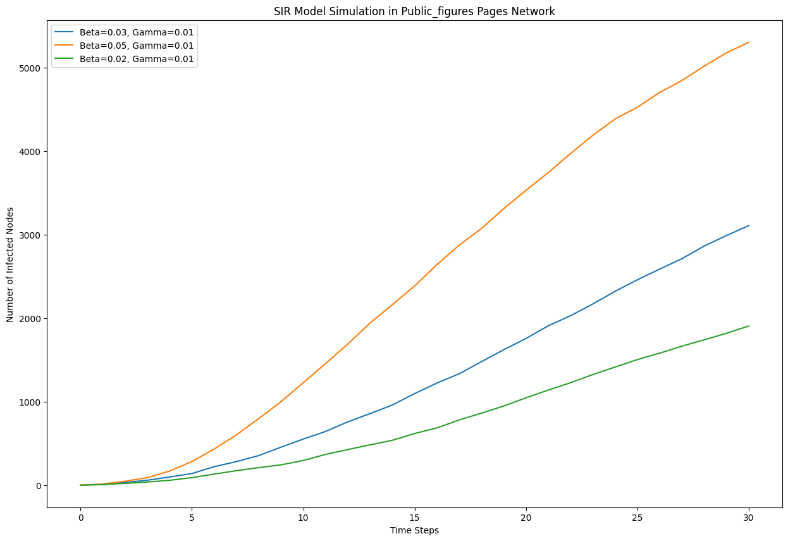
4.2

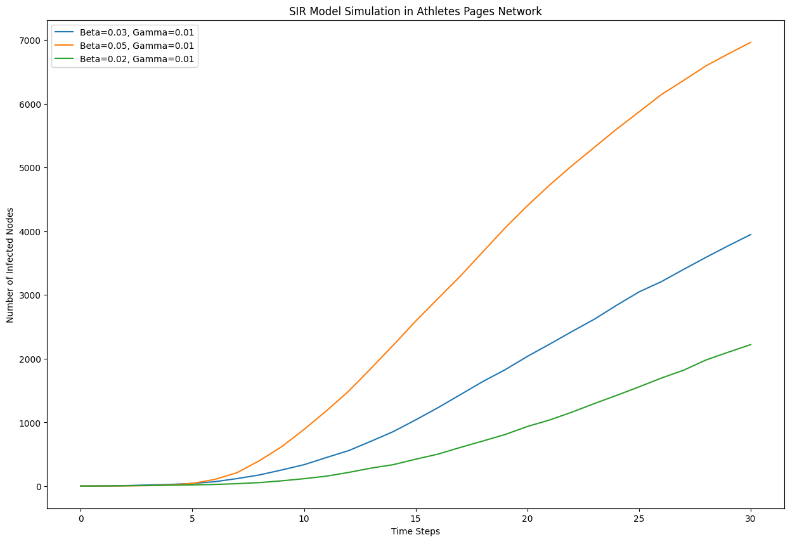
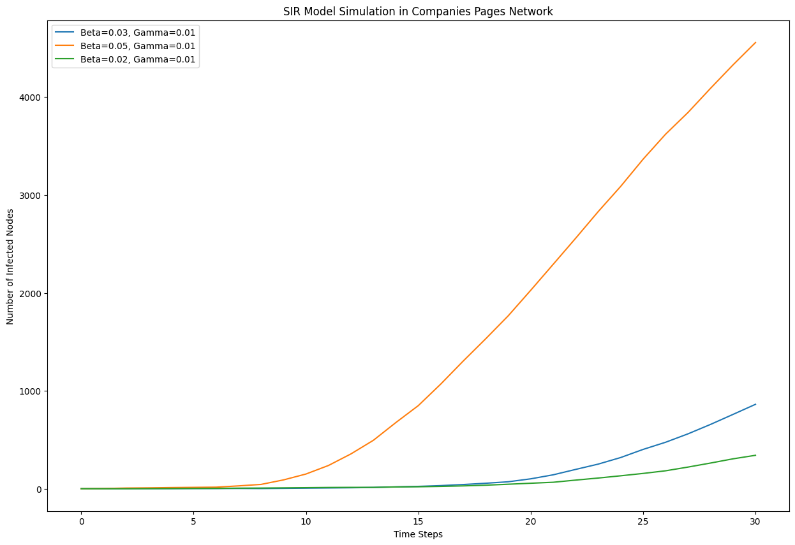
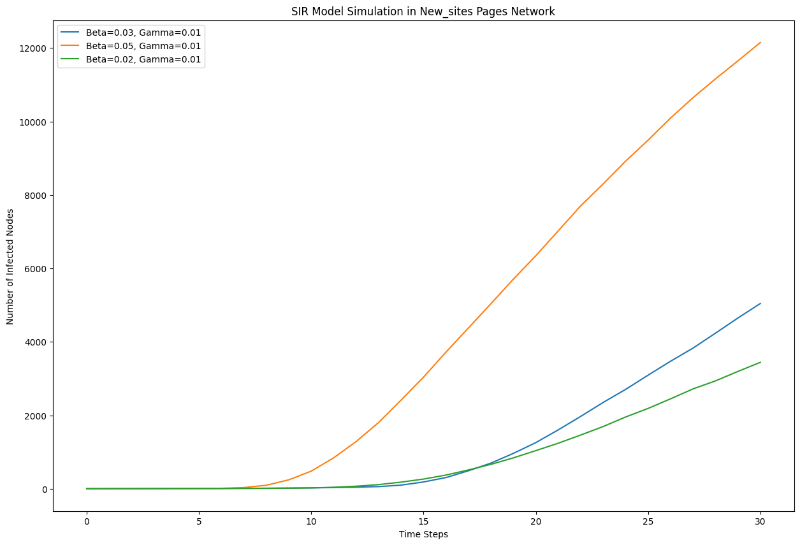
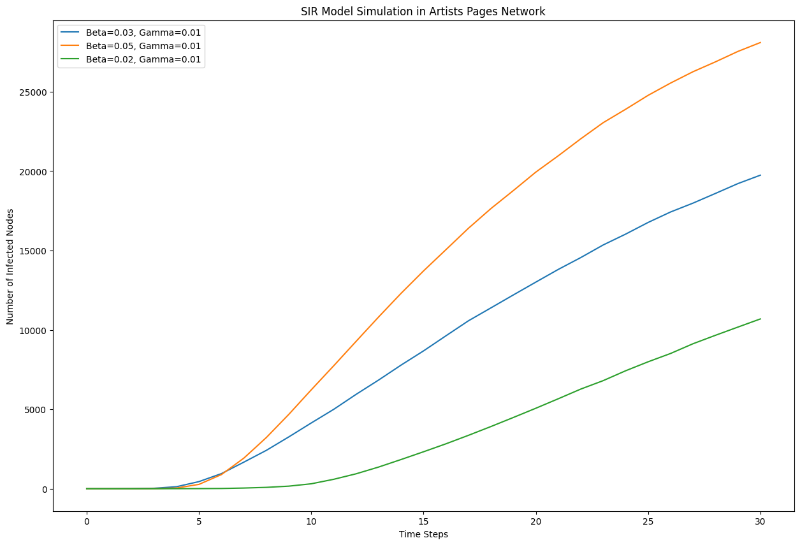
4.3

4.4 Diffusion pattern analysis of different types of users

The diffusion process of different types of information was simulated in the network. By setting initial infected nodes and propagation parameters (such as infection rate β and recovery rate γ), the dissemination process of information in the network was simulated. The diffusion range and speed of different types of information over certain time steps were recorded. The spread patterns of government, news sites, athletes, public figures, TV shows, politicians, artists, and companies information in the network were compared, identifying their similarities and differences in propagation range and speed. The experimental results were visually displayed to show the diffusion paths of different types of information.

Government information spreads at a moderate speed with a wide diffusion range, covering most network nodes within a short period. News sites information spreads quickly with a wide diffusion range, rapidly reaching various parts of the network. Athletes information spreads at a moderate speed with a concentrated diffusion range, primarily within specific communities. Public figures information spreads at a moderate speed with a wide diffusion range, showing high dissemination effects. TV shows information spreads at a moderate speed with a smaller diffusion range but effectively among specific nodes. Politicians information spreads quickly with a wide diffusion range, rapidly covering a large number of network nodes. Artists information spreads at a moderate speed with a wide diffusion range, maintaining dissemination over a longer period. Companies information spreads slowly with a relatively smaller diffusion range but has a longer duration of propagation.

The experimental results indicate significant differences in the spread patterns of different types of information in social networks. Government information spreads at a moderate speed initially, covering a wide range and then stabilizing gradually. News sites information, due to its timeliness and broad appeal, spreads quickly and covers the entire network rapidly. Athletes information primarily spreads within specific communities, with a concentrated diffusion range. Public figures information shows high dissemination effects and spreads widely. TV shows information spreads effectively among specific nodes but has a smaller overall diffusion range. Politicians information spreads quickly, covering a large number of network nodes rapidly. Artists information spreads at a moderate speed and maintains dissemination over a longer period. Companies information spreads slowly, with a smaller diffusion range but a longer duration of propagation.

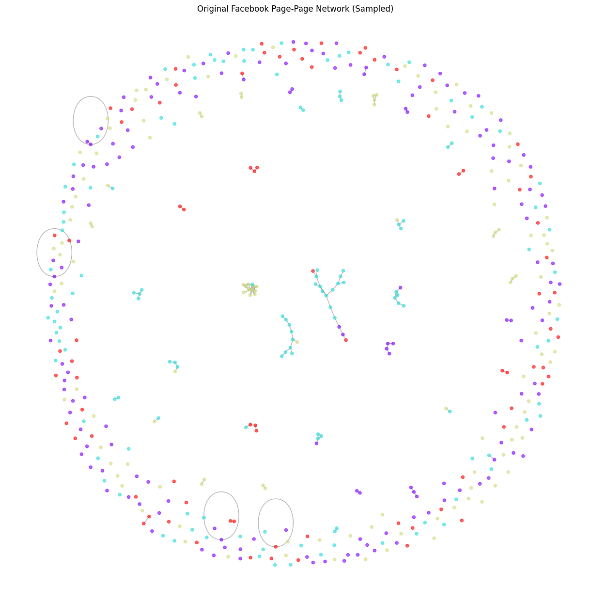
In each type of user network, we simulated the information propagation process. The initial infected nodes and propagation parameters, such as infection rate (β) and recovery rate (γ), were set to simulate the diffusion of information in the network. The diffusion range and speed of different types of information in a specific time step were recorded, and the diffusion paths and characteristics of different types of information were displayed through visualization methods.

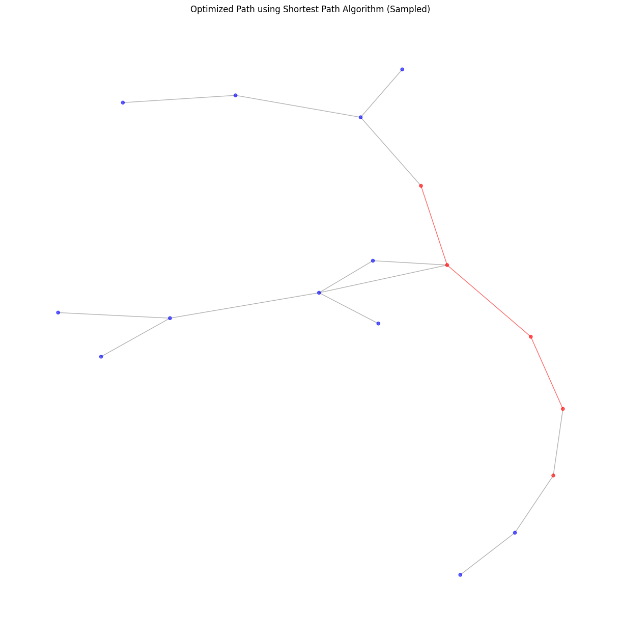
For the government user network, the simulation results show that the information diffusion of news, rumors, and advertisements has its own characteristics. The highest infection rate (β=0.05) leads to rapid information spread, covering most of the network nodes in a short time. Lower infection rates (β=0.03 and β=0.02) lead to slower propagation and a smaller number of infected nodes. News information spreads faster in the government network and quickly reaches a high number of infected nodes, while rumors spread the fastest and have the highest number of infected nodes. Advertising spreads relatively slowly, but can also cover a wider network for a longer time. In the news website user network, news information spreads the fastest, showing a wide range of spread and a high number of infected nodes. The information diffusion speed of rumors follows closely, quickly covering most of the nodes in the network. Advertising information spreads relatively slowly, but the number of infected nodes increases gradually over time.

For the athlete user network, news and rumors spread at a moderate speed. The highest infection rate causes information to spread rapidly within a specific community, while lower infection rates show a more concentrated spread pattern. Advertising information spreads slowly in this network, but eventually affects a considerable number of nodes. The public figure user network has a high information diffusion efficiency. News and rumors spread rapidly in this network, especially under the highest infection rate condition, the number of infected nodes increases rapidly. Advertising information spreads slowly in this network, but its spread range gradually expands over time.

In the TV program user network, information spreads slower than other types. Although information spreads relatively quickly under the highest infection rate, in general, the spread of news, rumors, and advertising is concentrated in specific nodes, showing the specificity and concentration of their audiences. For the politician user network, news and rumors spread quickly, especially under the highest infection rate condition, quickly covering most of the nodes of the network. Advertising information spreads slowly, but eventually affects a considerable number of nodes.

In the artist user network, the information diffusion of news and rumors shows high propagation efficiency. The information diffusion range under the highest infection rate is wide, and the number of infected nodes increases rapidly. Advertising information spreads slowly in this network, but eventually covers a large network range. Information propagation speed in the company user network is slow. Even under the highest infection rate conditions, the speed of information diffusion is slow, but over time, the information spread range of news, rumors, and advertisements gradually expands.

* 1. Optimization in Social Networks

Through the shortest path algorithm, we find the optimal path for information dissemination. We can see the result in Table 2. The results show that the optimized shortest path length is 5. This means that information propagated in the network can reach the target node in the shortest number of steps through this optimal path, which significantly reduces the time for information propagation. This optimization result shows significant improvement compared to the original network diameter 9. The shortest path algorithm improves the efficiency of information dissemination by reducing the propagation time. This result has important application value for scenarios where key information is expected to be disseminated in the shortest possible time, such as breaking news dissemination or emergency notification issuance.

On the other hand, the application results of the maximum flow algorithm show that the maximum flow value from the source node to the destination node is 1. This result shows that, given the network structure and constraints, the maximum flow of information or resources from source to destination is 1. Although the maximum flow value is low, by optimizing the path, the effective delivery of information in the network can be ensured. The maximum flow algorithm improves the stability and continuity of information transmission by optimizing the path of information flow in the network. This is of great significance for scenarios where stable transmission of information or resources in the network needs to be ensured, such as data synchronization or resource allocation.

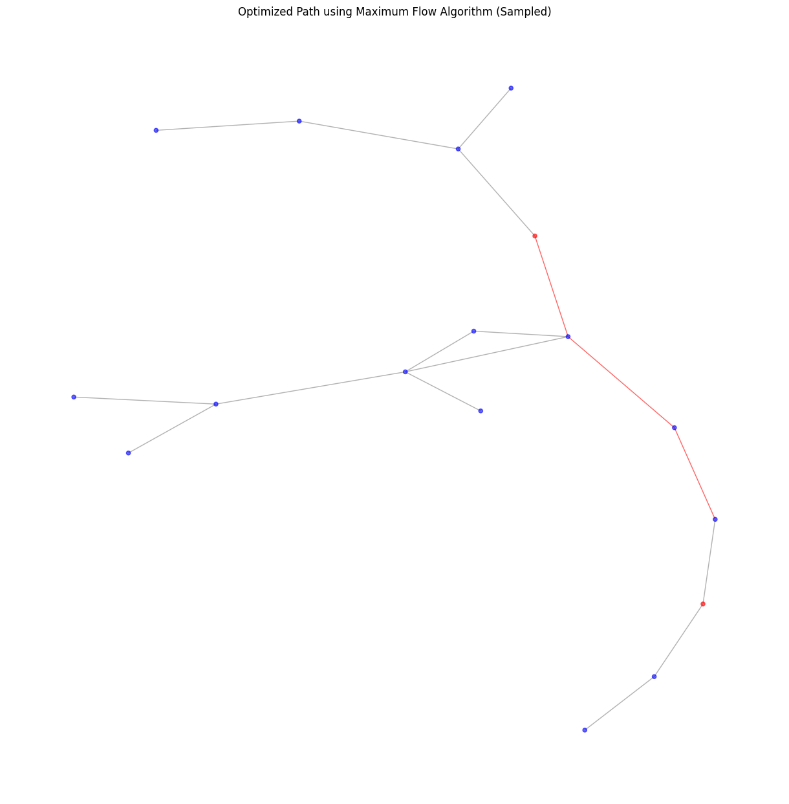


Table 2: Simulation Results

|  |  |
| --- | --- |
| Shortest Path Length | 5 |
| Maximum Flow Value | 1 |
| Original Network Diameter (Sampled) | 9 |
| Shortest Path Length after Optimization | 5 |