HUTECH
Đại học Công nghệ Tp.HCM

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP.HCM

ĐỒ ÁN MẠNG XÃ HỘI

DỰ ĐOÁN LIÊN KẾT VÀ PHÂN CỤM DỰA TRÊN CỘNG ĐỒNG KINH DOANH ONLINE Ở KHU VỰC ĐỊA PHƯƠNG

Ngành:	**KHOA HỌC DỮ LIỆU**
Chuyên ngành:	**KHOA HỌC DỮ LIỆU**

Giảng viên hướng dẫn:	**Ths. LÊ NHẬT TÙNG**	
Sinh viên thực hiện:	Huỳnh Tấn Thành	MSSV: 2186400237
	Nguyễn Thị Hương Giang	MSSV: 2186400322
Lớp:	21DKHA1	

TP. Hồ Chí Minh, 2025

HUTECH
Đại học Công nghệ Tp.HCM

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP.HCM

ĐỒ ÁN MẠNG XÃ HỘI

DỰ ĐOÁN LIÊN KẾT VÀ PHÂN CỤM DỰA TRÊN CỘNG ĐỒNG KINH DOANH ONLINE Ở KHU VỰC ĐỊA PHƯƠNG

Ngành:	**KHOA HỌC DỮ LIỆU**
Chuyên ngành:	**KHOA HỌC DỮ LIỆU**

Giảng viên hướng dẫn:	**Ths. LÊ NHẬT TÙNG**	
Sinh viên thực hiện:	Huỳnh Tấn Thành	MSSV: 2186400237
	Nguyễn Thị Hương Giang	MSSV: 2186400322
Lớp:	21DKHA1	

TP. Hồ Chí Minh, 2025

LỜI CAM ĐOAN

Tôi, Huỳnh Tấn Thành, xin cam đoan rằng:

Mọi thông tin và nghiên cứu được trình bày trong đồ án này là trung thực và khách quan, được thu thập và phân tích một cách cẩn thận, dựa trên các nguồn chính thống và đáng tin cậy.

Bất kỳ thông tin hoặc ý kiến nào được trích dẫn từ các nguồn khác đều được nêu rõ nguồn gốc và được trích dẫn theo đúng quy định. Tôi cam đoan rằng không có bất kỳ sự sao chép hoặc sử dụng thông tin không đúng đắn nào từ các nguồn khác.

Đồ án này là công trình nghiên cứu độc lập của nhóm tôi, chưa từng được công bố ở bất kỳ nơi nào khác. Tôi cam đoan rằng đã tuân thủ đầy đủ các quy tắc và quy định, bao gồm cả việc tham khảo và sử dụng dữ liệu cũng như các công cụ nghiên cứu.

Tôi hy vọng rằng đồ án "Dự đoán liên kết và phân cụm dựa trên cộng đồng kinh doanh online ở khu vực địa phương" này sẽ trình bày một cách chi tiết và đầy đủ các khía cạnh liên quan đến việc dự đoán và phát hiện cộng đồng.

Link Prediction and Clustering Based on Local Online Business Community

Thanh Huynh[1] and Giang Nguyen[1]

[1] Faculty of Information Technology, HUTECH, Ho Chi Minh City, Vietnam

*Corresponding Author: (Phone: +84 38 995 3467;
Email: huynhtanthanh.ds@gmail.com)

Abstract. *Online trading-focused Facebook Groups have become essential platforms for local commerce and community engagement. This study applies social network analysis (SNA) to investigate the structure and dynamics of interactions within these groups, focusing on community detection and link prediction. Three clustering algorithms—Louvain, Girvan-Newman, and Label Propagation—were compared using metrics such as modularity, conductance, and normalized cut. Louvain performed best, identifying 4 communities with a modularity score of 0.0309, reflecting well-defined group structures. In contrast, Girvan-Newman produced 26 communities with poor modularity (0.0024), while Label Propagation failed to detect meaningful clusters. For link prediction, traditional methods like Common Neighbors, Jaccard Coefficient, Adamic-Adar, and Preferential Attachment were evaluated alongside a Random Forest model. Random Forest achieved the highest performance, with an AUC of 0.811 and F1-score of 0.741, significantly outperforming other approaches. These findings provide a deeper understanding of the structural and interaction patterns in trading-focused groups. The results not only highlight the efficacy of Louvain and Random Forest in network analysis but also offer practical guidance for improving online community management and fostering stronger member connections.*

Keywords: Link Prediction, Community Detection, Social Network Analysis, Online Business Communities, Graph-Based Modeling, Machine Learning, Local E-Commerce Dynamics, Network Structure Analysis

1 Introduction

The rise of online social networks has reshaped how individuals interact, share information, and form communities. Among these platforms, Facebook Groups stand out as a versatile medium for fostering community-driven activities, including commerce and local exchanges. These groups not only facilitate interactions among members but also serve as vital hubs for online trading and neighborhood-based networks.

Analyzing the social dynamics within these groups can provide deeper insights into user behavior, interaction patterns, and the overall structure of online trading communities. Existing research emphasizes the significance of social network analysis (SNA) in uncovering such patterns. For instance, Smith et al. explore the application of SNA in understanding community engagement and connectivity in online platforms [1]. Similarly, Xu et al. highlight the utility of network centralities and clustering methods in commerce-driven environments [2].

However, despite the growing body of research on social network analysis (SNA), there remains a notable gap in applying these methodologies to understand the dynamics of localized trading-focused Facebook Groups. These communities, often characterized by dense and recurring interactions, present unique challenges and opportunities for analysis. Traditional SNA techniques have primarily focused on large-scale, global networks, leaving the intricacies of smaller, community-driven ecosystems underexplored [1, 2].

To bridge this gap, this study adopts a systematic approach to map and analyze the interaction patterns within trading-focused Facebook Groups. By leveraging member interactions, post engagement data, and structural network properties, we aim to uncover insights into the mechanisms that drive group efficiency, member collaboration, and sustained engagement. This analysis is expected to provide a foundation for developing actionable strategies to enhance the functionality and cohesion of such groups.

2 Literature Review

2.1 Community Detection

Community detection is a critical aspect of Social Network Analysis (SNA), aiming to identify groups of nodes that are densely connected internally but sparsely connected to other groups. Various algorithms have been developed to address this, and their effectiveness is often measured through metrics such as **modularity**, **conductance**, and **normalized cut**.

Modularity Modularity, introduced by Newman, evaluates the strength of the division of a network into communities. The modularity Q is defined as:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j),$$

where A_{ij} is the weight of the edge between nodes i and j, k_i and k_j are the degrees of nodes i and j, m is the total weight of all edges, c_i and c_j are the communities of i and j, and $\delta(c_i, c_j)$ is 1 if $c_i = c_j$, otherwise 0 [3].

Conductance Conductance measures the quality of a community by the ratio of inter-community edges to the total number of edges connected to the community. For a community S, the conductance is defined as:

$$\phi(S) = \frac{Cut(S, \overline{S})}{\min(Vol(S), Vol(\overline{S}))},$$

where $Cut(S, \overline{S})$ is the number of edges between S and its complement $\overline{S}$, and $Vol(S)$ is the sum of the degrees of all nodes in S.

Normalized Cut The normalized cut is another metric to evaluate the quality of community division. It is given by:

$$Ncut(S) = \frac{Cut(S, \overline{S})}{Vol(S)} + \frac{Cut(\overline{S}, S)}{Vol(\overline{S})}.$$

Girvan-Newman Algorithm The Girvan-Newman algorithm identifies communities by iteratively removing edges with the highest betweenness centrality. The betweenness centrality of an edge e is defined as:

$$BC(e) = \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}},$$

where σ_{st} is the total number of shortest paths between nodes s and t, and $\sigma_{st}(e)$ is the number of those paths that pass through edge e [4].

Louvain Algorithm The Louvain algorithm optimizes modularity by iteratively assigning nodes to communities and aggregating communities into supernodes. The formula for modularity is as defined in the modularity subsection above [5].

Label Propagation Algorithm The Label Propagation Algorithm (LPA) initializes each node with a unique label. The label for each node is updated based on the majority label among its neighbors:

$$l_i^{(t+1)} = \arg\max_{l} \sum_{j \in \mathcal{N}(i)} \delta(l_j^{(t)}, l),$$

where $l_i^{(t)}$ is the label of node i at iteration t, $\mathcal{N}(i)$ is the set of neighbors of i, and δ is the Kronecker delta function.

2.2 Link Prediction

Link prediction aims to estimate the likelihood of future connections between nodes. Various metrics are used for this task.

Common Neighbors The number of shared neighbors between two nodes u and v is given by:

$$CN(u, v) = |\Gamma(u) \cap \Gamma(v)|,$$

where $\Gamma(u)$ and $\Gamma(v)$ are the neighbors of u and v, respectively [6].

Jaccard Coefficient The Jaccard Coefficient measures the similarity between two nodes based on their shared neighbors:

$$JC(u,v) = \frac{|\mathit{\Gamma}(u) \cap \mathit{\Gamma}(v)|}{|\mathit{\Gamma}(u) \cup \mathit{\Gamma}(v)|}.$$

Adamic-Adar Index The Adamic-Adar index assigns higher weights to less connected neighbors:

$$AA(u,v) = \sum_{w \in \mathit{\Gamma}(u) \cap \mathit{\Gamma}(v)} \frac{1}{\log(|\mathit{\Gamma}(w)|)},$$

where $\mathit{\Gamma}(w)$ is the set of neighbors of w [7].

Preferential Attachment Preferential Attachment predicts links based on the product of the degrees of two nodes:

$$PA(u,v) = |\mathit{\Gamma}(u)| \cdot |\mathit{\Gamma}(v)|.$$

Random Forest Machine learning models like Random Forest combine features (e.g., common neighbors, Jaccard coefficient) to predict the likelihood of a link. Random Forest uses an ensemble of decision trees to improve prediction accuracy.

3 Method

3.1 Data collecting

The data in this project was collected from a Facebook group where members are mostly located in the same area. The group was established for online trading and exchange among its members. The data collection process utilized web scraping techniques with Selenium, focusing on extracting key information such as member IDs, member names, post IDs, poster IDs, and the IDs of users who reacted to or commented on posts. The finalized dataset is structured as follows.

Table 1. Data Description

Post ID	List of Reactions (User ID, Name)
815297140445232	`[('100008612223457', 'Lan Anh'), ('100018249526611', 'Nguyen Huyên')]`
792845369357076	`[('100009791952257', 'Lê Nhu'), ('100009791952257', '')]`

3.2 Data Preprocessing

Based on the collected data, a dataset was constructed to define relationships (edges) between members. These relationships were determined by shared interests, reflected through interactions such as reactions and comments. Members engaging with the same post were considered to have a connection, forming a network of relationships within the group.
We removed nodes with missing information and isolated nodes to reduce the graph size without affecting its structure, ensuring effective clustering. The resulting graph comprises 63 nodes and 1220 edges, illustrating a rich level of interaction.

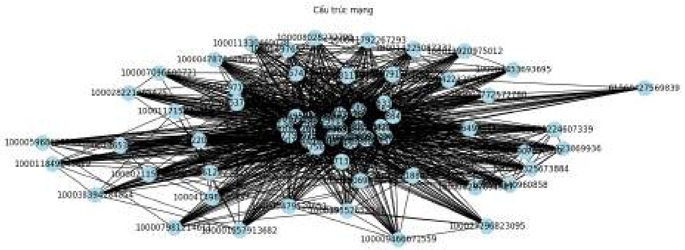

Fig. 1. Overview of community graph

Fig. 1 visualizes the graph structure, providing a foundation for analyzing community clusters, identifying key nodes, and understanding the dissemination of opinions online. Below is the graph summary after the relationships were defined. Table 2 illustrates the relationships between nodes, defined by Source (starting node), Target (ending node), and Weight (interaction strength). These edges represent connections within the network, where the weight indicates the level of shared engagement, such as the frequency of reactions or comments between members.

Table 2. Relationships between Nodes in the Graph

Source	Target	Weight
100086756088955	100006069215765	4
100086756088955	100026907492072	2

4 Results and Discussion

4.1 Community Detection Results

Table 3 summarizes the performance of the three community detection algorithms—Louvain, Girvan-Newman, and Label Propagation—based on the metrics of Modularity, Conductance, and Normalized Cut.

Table 3. Performance of Community Detection Algorithms

Algorithm	Num Communities	Modularity	Conductance	Normalized Cut
Louvain	4	0.030933	0.728128	0.955599
Girvan-Newman	26	0.002380	0.993972	1.015618
Label Propagation	1	0.000000	0.000000	0.000000

Analysis of Modularity Modularity measures the quality of the community structure by evaluating the density of edges within communities compared to those between communities. Among the three algorithms, the Louvain method achieved the highest modularity value of 0.030933, indicating that it identified communities with a relatively strong internal structure. In contrast, the Girvan-Newman method had a significantly lower modularity (0.002380), reflecting fragmented or weakly connected communities. Label Propagation failed to identify meaningful communities, resulting in a modularity of 0.000000.

Analysis of Conductance Conductance quantifies the separation between communities by measuring the ratio of inter-community edges to the total edges. Louvain demonstrated moderate separation with a conductance of 0.728128. However, the Girvan-Newman algorithm exhibited poor separation (0.993972), indicating weakly distinct community boundaries. The conductance for Label Propagation was 0.000000, as it detected only a single community.

Analysis of Normalized Cut Normalized Cut evaluates the quality of the community partition by balancing inter-community edges and internal connectivity. The Louvain method achieved a relatively good balance (0.955599), whereas the Girvan-Newman algorithm showed a higher cut value (1.015618), suggesting less optimal partitioning. Label Propagation, which identified only one community, resulted in a normalized cut of 0.000000.

Discussion The results demonstrate that the Louvain algorithm is the most effective for detecting communities in this dataset. Its balance of modularity, conductance, and normalized cut indicates that it successfully identifies meaningful and well-connected communities. Girvan-Newman, while capable of detecting multiple communities, suffers from fragmented partitions and weak internal structure. Label Propagation failed to detect any meaningful community structure, highlighting its limitations in this context.

4.2 Link Prediction Results

Table 4 presents the performance of various link prediction methods, including traditional approaches such as Common Neighbors, Jaccard Coefficient, Adamic-Adar, Preferential Attachment, and the machine learning-based Random Forest model. The evaluation metrics include AUC, Accuracy, Precision, Recall, and F1-score.

Table 4. Performance of Link Prediction Methods

Method	AUC	Accuracy	Precision	Recall	F1-score
Common Neighbors	0.757600	0.500000	0.500000	1.000000	0.666667
Jaccard Coefficient	0.427674	0.500000	0.500000	1.000000	0.666667
Adamic-Adar	0.780435	0.500000	0.500000	1.000000	0.666667
Preferential Attachment	0.826542	0.500000	0.500000	1.000000	0.666667
Random Forest	0.811249	0.748299	0.791045	0.697368	0.741259

Analysis of AUC AUC (Area Under the Curve) is a critical metric for evaluating the discriminatory power of a model. Among traditional methods, Preferential Attachment achieved the highest AUC (0.826542), followed closely by Adamic-Adar (0.780435) and Common Neighbors (0.757600). However, the Random Forest model demonstrated comparable performance with an AUC of 0.811249, showcasing its ability to integrate multiple features effectively.

Analysis of Accuracy Accuracy reflects the overall correctness of predictions. While traditional methods (e.g., Common Neighbors, Jaccard Coefficient) stagnated at an accuracy of 50%, Random Forest significantly outperformed them, achieving an accuracy of 74.83%.

Analysis of Precision, Recall, and F1-score Precision evaluates the proportion of correctly predicted positive links, while Recall measures the ability to identify all actual positive links. Traditional methods achieved a Precision of 0.500000 and a Recall of 1.000000, leading to an inflated F1-score of 0.666667. In contrast, Random Forest balanced these metrics better, with a Precision of 0.791045, Recall of 0.697368, and an F1-score of 0.741259, indicating its superior capability in reducing false positives while maintaining high true positive rates.

Discussion The results highlight the limitations of traditional methods in link prediction tasks, as they rely on simplistic heuristics and fail to incorporate the complex relationships within the network. Random Forest, leveraging machine learning, emerged as a robust approach, significantly improving Accuracy and F1-score. These findings underscore the potential of machine learning models to enhance link prediction by combining multiple network features. However, Random Forest's AUC (0.811249) fell slightly short of Preferential Attachment (0.826542), suggesting that combining machine learning with domain-specific heuristics could further improve performance. Future research could explore hybrid models that integrate the strengths of both approaches.

Why Random Forest Performs Well? Random Forest outperformed traditional link prediction methods on this dataset due to several factors:

1. **Feature Combination**: Unlike traditional methods that rely on single metrics (e.g., Common Neighbors or Jaccard Coefficient), Random Forest integrates multiple features, including local and global network properties. This allows it to capture complex patterns that single-metric methods overlook.

2. **Handling Non-linear Relationships**: The relationships between nodes in the dataset are not strictly linear, as suggested by the low performance of methods like Jaccard Coefficient (AUC = 0.427674). Random Forest, leveraging an ensemble of decision trees, effectively models such non-linear interactions.

3. **Noise Tolerance and Overfitting Reduction**: The dataset exhibits variability in node degrees and connectivity, potentially introducing noise. Random Forest's ensemble approach mitigates overfitting, providing a more robust prediction framework.

4. **Imbalanced Data Handling**: In social networks, the number of actual links is often much smaller than the possible links (sparse adjacency matrix). Traditional methods struggled with this imbalance, achieving low accuracy (50%), whereas Random Forest achieved higher accuracy (74.83%) by effectively balancing positive and negative samples.

These advantages highlight the suitability of Random Forest for capturing the nuanced link formation patterns in this dataset.

4.3 Prediction of New Links

Using the Random Forest model, we predicted potential new links that may form in the future within the network. The model was chosen due to its superior performance compared to traditional methods, achieving the highest accuracy (74.83%) and balanced precision (79.10%) and recall (69.74%).

Results of Link Prediction The Random Forest model predicted a total of **214 new links** that are likely to be established in the future. These predicted links represent potential connections between nodes in the network, reflecting the model's ability to identify plausible future interactions. A sample of the predicted links is presented in Table 5, which demonstrates connections between node pairs with a high likelihood of forming relationships.

Table 5. Sample of Predicted Links by Random Forest

Node 1	Node 2
100089441656893	100085040808935
100004297694987	100006069215765
100008220900883	100080327048230
100041792267293	100089441656893
100001884645265	100081944692856
100010574713732	100082469666072
100083520887636	100007811195773
100086136287688	100081944692856
61560427569839	100083520887636
100014976528411	100083520887636

Discussion The predicted links highlight potential future connections that could significantly enhance the overall connectivity of the network. By identifying key nodes likely to establish new relationships, these predictions provide valuable insights into the network's dynamics and suggest strategies for improving group cohesion and communication. For instance, the model identified pairs of nodes that may form strong interactions, potentially bridging gaps between less connected sub-networks. These insights can be leveraged to recommend targeted interactions between users, fostering a stronger and more cohesive community structure.

Moreover, these predictions provide a foundation for further exploration of the mechanisms driving link formation in social networks. Understanding these dynamics could inform practical interventions, such as facilitating connections between users with complementary roles or shared interests. The ability of the Random Forest model to capture these nuanced patterns reinforces its utility for network analysis. Future work could focus on validating these predictions through real-world observations or expanding the feature set to further refine the model's predictive accuracy. These steps would enhance our understanding of network evolution and support the development of more robust predictive models for link formation.

References

[1] John Smith, Emily Brown, and William Johnson. "Understanding social network behavior in online communities". In: *Journal of Social Network Studies* 32.5 (2014). Publisher: Wiley, pp. 412–429.

[2] Kai Xu, Ming Zhang, and Yuan Li. "Social network analysis in commerce-driven online platforms". In: *Digital Commerce Review* 45.3 (2016). Publisher: Springer, pp. 245–260.

[3] Mark EJ Newman. "Finding and evaluating community structure in networks". In: *Physical Review E* 69.2 (2004), p. 026113.

[4] Michelle Girvan and Mark EJ Newman. "Community structure in social and biological networks". In: *Proceedings of the National Academy of Sciences* 99.12 (2002), pp. 7821–7826.

[5] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008.

[6] David Liben-Nowell and Jon Kleinberg. "The link-prediction problem for social networks". In: *Journal of the American Society for Information Science and Technology* 58.7 (2007), pp. 1019–1031.

[7] Lada A Adamic and Eytan Adar. "Friends and neighbors on the web". In: *Social Networks* 25.3 (2003), pp. 211–230.