Pluralistic Homophily in Networks

Pontificia Universidad
JAVERIANA
Cali

Fernando Barraza Alvarado

Pontificia Universidad Javeriana Cali
Faculty of Engineering and Sciences
Cali, Colombia
2024

Pluralistic Homophily in Networks

Thesis submitted in partial fulfillment
of the requirements for the degree of:
DOCTOR OF ENGINEERING AND APPLIED SCIENCES

Fernando Barraza

Advisor:
Ph.D. Alejandro Fernández

Co-Advisor:
Ph.D. Carlos Ramirez

Pontificia Universidad Javeriana Cali
Faculty of Engineering and Sciences
Cali, Colombia
2024

Acknowledgments

Reflecting on the journey, I am deeply grateful for the support and inspiration of those who have accompanied me along the way:

To my advisor Alejandro Fernandez and my co-advisor Carlos Ramirez, for their invaluable guidance and support throughout these years. To the leadership of Pontificia Universidad Javeriana Cali, especially Andrés Jaramillo, for their unwavering support and understanding during the most challenging times. To all the professors of the Faculty of Engineering and Sciences, who shared their knowledge during classes and work sessions, where I got to know their human and professional qualities. Special thanks to Gloria Álvarez for her always accurate judgment, and to Abel Álvarez for his mathematical contributions. Also, thanks to Jorge Finke for inspiring me and igniting my passion for complex networks. To my cohort mates, with whom we navigated the path together and can now proudly say we have made it. To Universidad de San Buenaventura Cali, for giving me the opportunity to start these studies and for being flexible to help me achieve my goal. To all my friends, with whom I could discuss aspects of my work or who were simply there for me. And of course, to all my family and my wife's family, whom I consider my own. Especially to my mother and sisters, always in the background and attentive to my progress. To my wife Yanira and my children Carlos Andrés and Salomé. Their constant support, trust, and patience have been a driving force throughout this time. There were many hours and moments we could not share, but they understood that it was all for a common good. I love you. To my father (R.I.P.), because I always know you are there, as when in my dreams you adviced me to have into account the sign change in the QR decomposition. It was brilliant!! And finally to God, how I live and feel you.

Contents

List of Figures

List of Tables

Resumen

El análisis de redes ha revelado patrones importantes de homofilia, también conocida como asortatividad o mezcla asortativa, donde los nodos tienden a vincularse con nodos similares. Este fenómeno es especialmente notable en las redes sociales, en contraste con las redes biológicas y tecnológicas. Sin embargo, las medidas tradicionales de asortatividad no capturan completamente la complejidad de las interacciones en redes con comunidades superpuestas, donde los nodos pueden pertenecer a múltiples comunidades simultáneamente.

El concepto de homofilia pluralista fue introducido por otros autores para describir la tendencia de los nodos a conectarse con otros que comparten membresías comunitarias mutuas. La homofilia pluralista destaca cómo los nodos, particularmente en las estructuras de núcleo-periferia, frecuentemente comparten múltiples membresías con sus vecinos, lo que lleva a un alto grado de similitud bajo uno o más atributos comunes. Motivada por estas observaciones, esta tesis realiza un estudio profundo de la homofilia pluralista, desarrollando una perspectiva holística del fenómeno observado por autores anteriores, que incluye dos métricas propuestas en este trabajo, complementadas por un conjunto de métricas bien conocidas en el campo de la teoría de redes complejas. La primera métrica caracteriza el fenómeno de la homofilia pluralista a nivel de red, utilizando el coeficiente de correlación de Pearson aplicado al número de membresías comunitarias de los nodos, es decir, al número de comunidades a las que pertenecen los nodos de la red. La segunda métrica caracteriza el fenómeno a nivel local, considerando las interacciones de los nodos individuales dentro de la red, es decir, los enlaces con sus nodos vecinos. Aunque estas métricas no miden directamente la homofilia pluralista tal como fue observada originalmente, es decir, la tendencia de los nodos a conectarse con otros que comparten sus mismas comunidades, tienen el potencial de extenderse para lograrlo. Este enfoque más holístico permite abordar el estudio de la homofilia pluralista en sus diferentes variantes y contextos, mostrando su versatilidad y aplicabilidad en diversos campos.

El análisis empírico de nueve redes distintas, incluidas redes sociales, de colaboración y biológicas, reveló que cuatro de ellas exhibían altos niveles de homofilia pluralista en términos de membresías comunitarias de los nodos, mientras que las redes restantes mostraban patrones no homofílicos. Además, se observó que los nodos con menos membresías comunitarias tienden a exhibir una mayor homofilia pluralista que el promedio de la red, mientras que los nodos con alta membresía se conectan preferentemente tanto dentro como entre comunidades, mostrando una gama más amplia de interacciones pluralistas. Además, los nodos de bajo grado generalmente no son homofílicos, y su homofilia pluralista se vuelve más pronunciada—ya sea positiva o negativamente—a medida que aumenta el grado del nodo.

Estos hallazgos llevaron al desarrollo de un marco analítico que permite una comprensión más profunda del comportamiento pluralista en redes con comunidades superpuestas. Este marco no solo aclara la naturaleza de la homofilia pluralista, sino que también proporciona nuevas herramientas para la predicción de atributos de nodos y otras tareas analíticas en el campo de la ciencia de redes.

El impacto de esta investigación radica en su capacidad para proporcionar una medida más precisa de las interacciones complejas en redes con comunidades superpuestas, superando las limitaciones de las métricas tradicionales de homofilia. En consecuencia, esto tiene implicaciones significativas para el análisis de redes sociales y de colaboración, facilitando una mejor comprensión de los patrones de conectividad y las dinámicas subyacentes.

Abstract

Network analysis has revealed important patterns of homophily, also known as assortativity or assortative mixing, where nodes tend to link with similar nodes. This phenomenon is particularly notable in social networks, in contrast to biological and technological networks. However, traditional measures of assortativity do not fully capture the complexity of interactions in networks with overlapping communities, where nodes can belong to multiple communities simultaneously.

The concept of pluralistic homophily was introduced by other authors to describe the tendency of nodes to connect with others that share mutual community memberships. Pluralistic homophily highlights how nodes, particularly in core-periphery structures, frequently share multiple memberships with their neighbors, leading to a high degree of similarity under one or more common attributes. Motivated by these observations, this thesis conducts an in-depth study of pluralistic homophily, developing a holistic perspective of the phenomenon observed by previous authors, which includes two metrics proposed in this work, complemented by a set of well-known metrics in the field of complex network theory. The first metric characterizes the phenomenon of pluralistic homophily at the network level, using the Pearson correlation coefficient applied to the number of community memberships of the nodes, that is, the number of communities to which the nodes in the network belong. The second metric characterizes the phenomenon at the local level, considering the interactions of individual nodes within the network, that is, the links with their neighboring nodes. Although these metrics do not directly measure pluralistic homophily as originally observed, that is, the tendency of network nodes to connect with nodes that share their same communities, they have the potential to be extended to do so. This more holistic approach allows addressing the study of pluralistic homophily in

its different variants and contexts, demonstrating its versatility and applicability in various fields.

The empirical analysis of nine distinct networks, including social, collaborative, and biological networks, revealed that four of them exhibited high levels of pluralistic homophily in terms of the community memberships of the nodes, while the remaining networks showed non-homophilic patterns. Additionally, it was observed that nodes with fewer community memberships tend to exhibit greater pluralistic homophily than the network average, while nodes with high membership preferentially connect both within and between communities, showing a broader range of pluralistic interactions. Furthermore, low-degree nodes are generally non-homophilic, and their pluralistic homophily becomes more pronounced—whether positively or negatively—as the node degree increases.

These findings led to the development of an analytical framework that allows a deeper understanding of pluralistic behavior in networks with overlapping communities. This framework not only clarifies the nature of pluralistic homophily but also provides new tools for predicting node attributes and other analytical tasks in the field of network science.

The impact of this research lies in its ability to provide a more accurate measure of complex interactions in networks with overlapping communities, overcoming the limitations of traditional homophily metrics. Consequently, this has significant implications for the analysis of social and collaborative networks, facilitating a better understanding of connectivity patterns and underlying dynamics.

1 Introduction

Homophily, known as *assortativity* or *assortative mixing*, is a fundamental concept in network analysis. It describes the tendency of nodes to link with similar nodes, where similarity can be based on characteristics such as node degree [1]. Social networks often exhibit high assortativity, in contrast to biological and technological networks, which tend to show low assortativity, or *disassortativity* [2]. It is important to note that, generally, homophily refers to positive assortativity, that is, the tendency of nodes to connect with similar nodes. The term "homo" refers to similarity. However, in this thesis, the term homophily will be used in a general sense to refer to the phenomenon of assortativity, whether positive or negative.

Communities in network science are groups of nodes with a higher likelihood of connecting to each other than to nodes from other communities [3]. Community detection techniques include differential equations, random walks, spectral clustering, and modularity maximization. Networks can have disjoint communities or overlapping communities, where some nodes belong to multiple communities. The research by Yang and Leskovec [4] found that overlapping community areas have higher connection densities compared to non-overlapping areas. They introduced the concept of *pluralistic homophily*, where nodes tend to connect with others sharing mutual community memberships.

Building on this concept, this thesis proposes an approach to measure *pluralistic homophily*, extending the traditional concept of assortativity to better capture the tendency of nodes to connect with others having a similar number of community memberships. The approach includes metrics to conduct analysis both at the network-level and the at local-level. A first metric, based on the concept of assortativity [1], measures the tendency of nodes to connect with others that have a similar number of community memberships at the network level. A second metric, derived from the first and supported by the works of [5] and [6], quantifies the tendency of nodes to connect with others that have a similar number of community memberships but at the local level, focusing on individual nodes. These two novel metrics offer a more comprehensive perspective on pluralistic homophily.

The proposed measure is designed to be broadly applicable because it starts from the concept of node similarity based on the number of memberships they have, regardless of the specific communities to which they belong. This approach captures a wide range of homophilic behaviors, reflecting the idea that nodes connect based on similarity in the number of communities to which they belong, rather than specific communities. In this sense, the phenomenon described by Yang and Leskovec, where nodes with shared community memberships are more likely to connect, could be viewed as a specific scenario within this broader conceptualization of pluralistic

homophily. In the theoretical background, we provide an outline of how this specific phenomenon could be examined by a simple specialization of these metrics, suggesting a potential avenue for further exploration in future work, demonstrating the flexibility of our approach.

To provide a clear understanding of this phenomenon, empirical analysis of nine distinct networks was conducted, revealing diverse patterns of pluralistic homophily. Four networks exhibited positive assortativity for community memberships of nodes (i.e., the tendency of nodes to connect with other nodes that have a similar number of community memberships), while others showed non-assortative patterns. Interestingly, three networks also exhibited degree-based disassortativity. Contrary to general trends, certain networks displayed opposing tendencies at the node level. Nodes with fewer community memberships often showed higher pluralistic homophily than the network average, whereas nodes with a higher number of community memberships preferred connections to nodes both with high and with low community memberships, indicating a broader range of interactions. Additionally, low-degree nodes generally exhibited non-assortative behavior, with pronounced pluralistic homophily as node degree increased. These findings led to the development of an analytical framework for a deeper understanding of pluralistic homophily in networks.

Understanding homophily and community structure is crucial for comprehending the dynamics and organization of different kinds of networks. This research aims to enhance this understanding by expanding the analysis of pluralistic homophily and introducing new elements into an analytical framework, thus improving the insight into network behavior and node interactions in overlapping communities.

By providing a more universal method to analyze pluralistic homophily, this work not only captures the specific homophilic tendencies observed in previous studies but also opens new perspectives for understanding the dynamics of community structures and node interactions in complex networks.

1.1 Problem Statement

The central problem addressed in this research is the comprehensive understanding and quantification of the phenomenon of pluralistic homophily in networks with overlapping communities. Traditional measures of assortativity do not fully capture the complexity of node interactions in such networks, where nodes often belong to multiple communities simultaneously. Resolving this complexity requires not only identifying and describing the phenomenon but also developing theoretical frameworks and analytical tools to measure and interpret it accurately. This research aims to develop and validate new metrics that can accurately measure pluralistic homophily at both the global and local levels, thus providing deeper insights into

connectivity patterns within social, collaborative and biological networks. In achieving this, the research seeks to advance theoretical and practical knowledge of network behavior and community structure.

1.2 Background

In this section, the foundational concepts relevant to the study are presented, including the definitions and formalism of networks and graphs, the concept of assortativity, local assortativity, community detection, and overlapping communities.

In this work, the term *network* is used to refer to the practical application of the mathematical concept of a *graph*.

Definition 1. A *network* is an ordered pair $G = (V_G, E_G)$, where V_G represents the set of vertices (or nodes) of the graph, with each vertex denoted by v, and E_G represents the set of edges (or links between vertices) of the graph, with each edge denoted by e. Formally, this can be defined as $V(G) \equiv V_G = \{v_1, v_2, \ldots, v_n\}$ and $E(G) \equiv E_G = \{e = (v_x, v_y) \mid v_x, v_y \in V_G\}$, where each edge e connects a pair of vertices (v_x, v_y) within the set of vertex V_G.

With the formal definition of networks and graphs established, the concept of assortativity, a key metric for analyzing the structure and characteristics of networks, is discussed next.

1.2.1 Assortativity

Assortativity, or assortative mixing, is the tendency of nodes in a network to associate with other nodes that share similar characteristics. This pattern of connectivity is often seen in social networks, where nodes (individuals) frequently create links based on similar characteristics such as interests, social position, or demographics [2]. On the other hand, other networks, like metabolic networks in biology, show a disassortative trend, meaning that nodes are more likely to connect with other nodes that have different characteristics [7]. Assortativity is a quantitative metric that evaluates the probability of nodes connecting with one another according to similar attributes. This metric produces a positive value in assortative networks, suggesting a tendency for homophily—the formation of links between nodes that have comparable characteristics.

The attributes of nodes within a network can be classified into two main categories: enumerative and scalar. Enumerative attributes are defined by a discrete and limited set of possible values. For example, within a social network, attributes such as gender, nationality, or profession fall into this category, where each individual (node) is associated with one specific category from a predefined list. In contrast,

scalar attributes are characterized by a continuous range of values, facilitating quantitative comparisons.

A common form of assortative mixing related to scalar attributes is degree mixing. In such instances, nodes with a high number of connections (high degree) are more likely to form links with other nodes that are similarly well-connected. Conversely, nodes with fewer connections (low degree) tend to associate with other nodes that also have a limited number of connections. This common form of assortative mixing, degree assortativity, is particularly prominent in the literature. Degree assortativity is extensively explored and studied within the realm of assortativity, highlighting the tendency for nodes with a large number of connections (high degree) to link with other high-degree nodes, while nodes with fewer connections (low degree) are more likely to associate with others that also have few connections.

Building on this foundational concept, the assortativity coefficient r quantifies the tendency of the nodes within a network to connect with others based on similar scalar attributes. This concept extends beyond the degrees of the nodes to encompass any scalar property of the nodes in V_G, facilitating a broader analysis of the structure of the network. The assortativity coefficient r is based on the Pearson correlation coefficient, which measures the linear correlation between two variables, providing a value between -1 and 1.

As defined in [1], the formula for calculating r is as follows:

$$r = \frac{\sum_{xy}(xy(e_{xy} - q_x q_y))}{\sigma_q^2} \tag{1}$$

where σ_q^2 represents the variance of the distribution q, indicating the normalized distribution of the scalar attributes between the nodes in V_G, e_{xy} denotes the joint probability distribution for pairs of vertices x and y, capturing the likelihood that there is an edge between the nodes with specific scalar attributes. q_x and q_y are the distributions of the scalar attributes for the two vertices at either end of a randomly chosen edge. The summation $\sum_{xy}$ covers all possible pairs of scalar attributes within V_G. As with the Pearson coefficient, the value of r ranges from -1 to 1, where $r = 1$ means perfect assortativity, $r = -1$ indicates perfect disassortativity, and $r = 0$ reflects a lack of assortative mixing, which implies random connectivity between nodes in V_G.

This measure allows for the assessment of how nodes with similar or dissimilar attributes connect within a network, providing valuable insights into the underlying structural tendencies and the potential influences of various nodal characteristics on network formation.

Figure 1 illustrates the concept of assortativity in networks. In Figure 1(a), an assortative network is shown where nodes with similar degrees are connected, re-

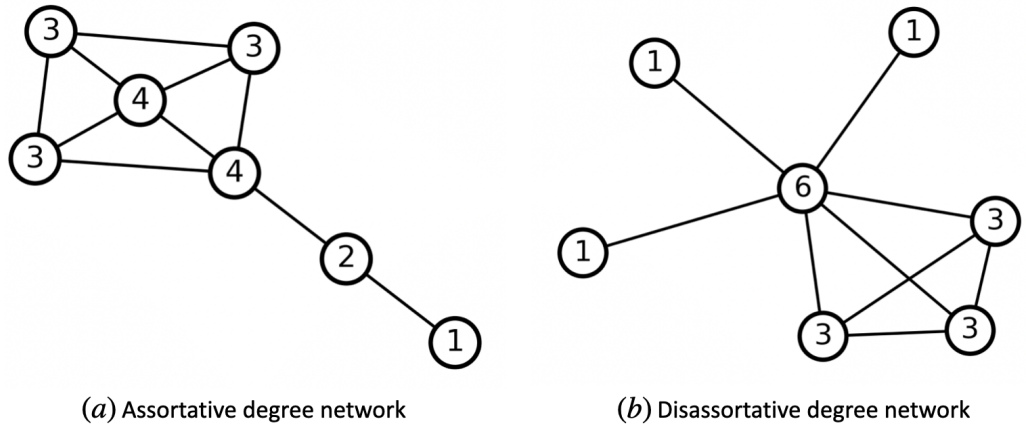


Figure 1: Assortativity for two networks. Each node is labeled with its degree. (a) Shows an assortative network with $r = 0.24$, where nodes are connected to others of similar degrees. (b) Shows a disassortative network with $r = -0.76$, where high-degree nodes are attached to low-degree nodes.

sulting in a positive assortativity coefficient ($r = 0.24$). Conversely, Figure 1(b) illustrates a disassortative network where high-degree nodes connect to low-degree nodes, leading to a negative assortativity coefficient ($r = -0.76$). These examples demonstrate how assortativity can provide insights into the structure and connectivity patterns within different types of networks.

Despite the robust framework provided by the assortativity coefficient, it is essential to consider various derivations of the main assortativity equation to measure assortativity according to different network types and configurations. Here are several alternative measures of assortativity:

Assortativity based on scalar attributes: The formula for assortativity can be generalized to any scalar property of a node, facilitating the analysis of networks based on various node attributes. The assortativity r, based on a scalar variable, is given by:

$$r = \frac{\sum_{xy} xy(e_{xy} - a_x b_y)}{\sigma_a \sigma_b} \tag{2}$$

where σ_a and σ_b are the standard deviations of the distributions a_x and b_y, respectively. e_{xy} is the probability of an edge connecting two nodes with the values x and y for the scalar property considered.

Assortativity based on discrete attributes: Assortativity can also be calculated based on any discrete attribute of the node. For example, in a social network, the node property 'gender' is an enumerated vertex characteristic. The assortativity of this network is given by:

$$r = \frac{\sum_i (e_{ii} - a_i b_i)}{1 - \sum_i a_i b_i} \tag{3}$$

where e is the matrix whose elements are e_{ij}, and a_i and b_i are the fractions of each type of end of an edge that is attached to vertices of type i.

Node-based network assortativity in directed networks: In directed networks, the directionality of edges requires a different approach to measure assortativity. The assortativity coefficient for out-degrees is given by:

$$r_{out} = \frac{1}{\sigma_{out,q}\sigma_{out,q'}} \left[\sum_{jk} jk e_{out,jk} - \mu_{out,q}\mu_{out,q'} \right] \tag{4}$$

where $\sigma_{out,q}$ and $\sigma_{out,q'}$ are the standard deviations of q_{out} and q'_{out} of the network, respectively.

Similarly, the assortativity coefficient for in-degrees is given by:

$$r_{in} = \frac{1}{\sigma_{in,q}\sigma_{in,q'}} \left[\sum_{jk} jk e_{in,jk} - \mu_{in,q}\mu_{in,q'} \right] \tag{5}$$

These measures allow for the evaluation of assortativity based on the directionality of links, providing insights into the connectivity patterns of directed networks.

Link-based local assortativity in undirected networks: The assortativity of a link in an undirected network can be measured using the following formula:

$$\rho_e = \frac{1}{M\sigma_q^2} \left[(j_i - \mu_q)(k_i - \mu_q) \right] \tag{6}$$

where M is the number of links, μ_q is the mean of the degree distribution, and σ_q is the standard deviation of the degree distribution. This measure evaluates the contribution of individual links to the overall network's assortativity.

By incorporating these diverse measures, a comprehensive understanding of assortativity in networks is obtained, addressing various attributes and configurations, and providing a nuanced view of the connectivity patterns within different types of networks.

1.2.2 Local Assortativity

In any type of network, certain nodes can establish connections in a way that follows the prevailing trend of the overall assortativity of the network. Even in networks characterized by generally positive or negative assortativity, it may occur that specific nodes show contrasting connection tendencies. Addressing these outliers, sev-

eral authors have introduced the notion of local assortativity, which evaluates the link tendencies of individual nodes in relation to the general trend of the network.

According to [8], the local assortativity of a node v can be quantified using a local assortativity coefficient r_v, which is defined as follows:

$$r_v = \frac{1}{\sigma_q^2} \sum_{i=1}^{d_v} (m_v - \mu_q)(m_i - \mu_q) \tag{7}$$

where d_v is the degree of node v, m_v is the attribute value of node v, m_i is the attribute value of node i, a neighbor of v, μ_q is the mean of the attribute values in the network, and σ_q^2 is the variance of the attribute values in the network. This coefficient measures the tendency of a node to form connections with others that have similar attributes.

The local assortativity coefficient r_v is derived as follows:

$$r_v = \frac{1}{\sigma_q^2} \sum_{i=1}^{d_v} (m_v - \mu_q)(m_i - \mu_q)$$

This formulation evaluates the assortativity of individual nodes in relation to the general trend of the network.

Subsequently, the same authors presented a new equation to measure local assortativity that eliminates the bias present in the previous formulation [9]. The initial formulation had a bias towards nodes with a low degree, which skewed the local assortativity measure. The unbiased local assortativity coefficient can be expressed as follows:

$$\rho_v = \frac{\alpha_v - \beta_v}{\sigma_q^2} \tag{8}$$

where

$$\alpha_v = \sum_{i=1}^{d_v} (m_v - \mu_q)(m_i - \mu_q)$$

and

$$\beta_v = \frac{(j+1)j\mu_q}{2M}$$

with j being the excess degree of node v, μ_q the mean of the attribute values in the network, σ_q^2 the variance of the attribute values in the network, and M the total number of edges in the network. This new coefficient corrects the bias by adjusting the contribution of each node to the network's assortativity.

The corrected local assortativity coefficient ρ_v is derived as follows:

$$\alpha_v = \sum_{i=1}^{d_v} (m_v - \mu_q)(m_i - \mu_q)$$

$$\beta_v = \frac{(j+1)j\mu_q}{2M}$$

The corrected local assortativity coefficient ρ_v is then given by:

$$\rho_v = \frac{\alpha_v - \beta_v}{\sigma_q^2}$$

This formulation ensures that the contribution of each node to the overall assortativity is properly scaled, eliminating the bias present in the previous formulation.

Despite the adjustment made in the previous equation, previous methods for measuring local assortativity still have some limitations. For instance, in the undirected case, if the average neighbor degree is higher than the expected degree, the node is considered assortative, and vice versa. This leads to peripheral nodes connected to similar peripheral nodes being considered disassortative, while hub nodes in a rich club are considered assortative despite variations in degrees among rich club members. This approach contradicts the global definition of assortativity.

To further improve upon these limitations, an alternative approach proposes a new measure for node-assortativity based on a node's contribution to network assortativity [10]. The derivation of this measure starts with the degree assortativity of a network, defined as the Pearson correlation coefficient:

$$r = \frac{1}{\sigma_q^2}\left(\sum_{jk} jke_{jk} - \mu_q^2\right) \tag{9}$$

where e_{jk} is the joint probability distribution of the degrees of the two nodes at either end of a randomly chosen link, q_k is the expected degree distribution, μ_q is the mean of q_k, and σ_q is its standard deviation. Using this definition, the contribution of an individual edge to the network assortativity can be expressed as:

$$r_e = \frac{1}{\sigma_q^2}\left((j-\mu_q)(k-\mu_q)\right) \tag{10}$$

In an undirected network, the local assortativity for a node v can be derived by summing the contributions of all edges connected to v, scaling by 0.5 to account for each edge being connected to two nodes:

$$r_v = \frac{M^{-1}}{2\sigma_q^2}\sum_{i=1}^{d_v}(j_i-\mu_q)(k_i-\mu_q) \tag{11}$$

Here, d_v is the degree of node v, j_i is the degree of node v, and k_i is the degree of the node at the other end of the i-th link. The term M appears in the numerator to normalize the sum by the total number of edges, ensuring that the contributions of all edges are properly accounted for. This measure provides a refined perspective

on local assortativity that aligns more closely with the global definition.

If we consider non-degree based assortativity, then assortativity is computed based on some scalar value which is a property of the node. In this case, the node assortativity of the node will be given by:

$$r_v = \frac{M^{-1}}{2\sigma_q^2} \sum_{i=1}^{d_v} (a_i - \mu_q)(b_i - \mu_q) \tag{12}$$

where, for each link i, a_i is the scalar value of the node at one end of the link and b_i is the scalar value of the node at the other end. The degree of node v is still d_v. Here, clearly $a_i \neq d_v$ and $b_i \neq d_v$ in general. The quantities μ_q y σ_q will also need to be calculated in terms of the scalar value considered.

The last measure overcomes the limitations of previous methods in several ways. Unlike earlier methods that might incorrectly classify peripheral or hub nodes, this measure evaluates the specific contribution of each node's connections to the overall network assortativity, thus providing a more accurate assessment. By focusing on the contribution of individual nodes, the new measure eliminates the bias towards nodes with low degrees, resulting in a more balanced and precise evaluation. Additionally, this approach ensures that the local measure is consistent with the global definition of assortativity, avoiding contradictions and providing a coherent understanding of the network structure.

1.2.3 Overlapping Communities

In network science, a community is defined as "a group of nodes that have a higher likelihood of connecting to each other than to nodes from other communities" [3]. The structure of communities within a network is rigorously defined.

Definition 2. A *community* is a subset of nodes, denoted as $c \subseteq V_G$, where V_G is the set of all nodes in the graph G. The detected communities on a graph form a set $C = \{c_1, c_2, \ldots, c_n\}$ such that each c_i is a subset of V_G (i.e., a family of sets where each set is in turn a set of vertices of V_G).

Communities can often be identified as ground-truth based on known characteristics or external information. However, in cases where ground-truth communities are not available, it becomes necessary to identify communities based on the structure of the network itself. In such cases, community detection techniques are essential for understanding the structure and function of networks. Various approaches, such as graph partitioning, hierarchical clustering, partitional clustering, and spectral clustering, are used to identify communities within networks [11]. These methods help in

uncovering the modular structure of networks and understanding the relationships between nodes.

Although traditional community detection methods focus on disjoint communities, recent research has highlighted the importance of considering *Overlapping Communities*, where nodes can belong to multiple communities simultaneously.

Definition 3. *Overlapping communities* are defined as communities where their intersection is non-empty, that is, $c_x \cap c_y \neq \varnothing$.

Overlapping communities are particularly relevant in social networks, where individuals often participate in multiple social groups. For example, in a social network like Facebook, an individual may be part of different overlapping communities such as family, friends, colleagues, and hobby groups. Although this phenomenon is more common in social networks, other types of networks also experience overlapping communities, which makes the study of these communities in social networks beneficial for understanding similar structures in other domains.

To detect overlapping communities, several prominent algorithms were developed early on, including Hierarchical Link Clustering (HLC) [12], Clique Percolation [13], Greedy Modularity [14], and Infomap [15]. These initial algorithms laid the foundation for the field, and numerous new proposals have emerged since, building on and refining these early approaches [16]. This thesis puts focus on HLC due to its unique approach and relevance to the topics discussed.

Hierarchical Link Clustering (HLC): Unlike traditional community detection methods based on comparing the similarity of the nodes, HLC makes a comparison of similarity between links. Links are arranged in a dendrogram based on their similarity that is calculated with a *Jaccard index* (or *Tanimoto index* for weighted similarities). The dendrogram is cut at different thresholds (from bottom to top) until the average density for all resulting partitions of links gives the maximum value. The communities are formed, one for every given partition of links. As each partition of links has correspondence with a single community, the memberships of the nodes to the communities are assigned depending on what nodes participate in each link. In this way, a node can be overlapped in as many communities as it has links in different partitions. An example illustrating how the HLC algorithm works will be presented in Section 5.3.

Evaluation Metrics for Overlapping Community Detection: Evaluating the quality of detected communities is crucial for understanding their relevance and effectiveness. Metrics such as modularity, conductance, and coverage are commonly

used to assess the strength and coherence of communities. Additionally, the resolution limit problem, which refers to the inability of some methods to detect smaller communities within large networks, is an important consideration when choosing a community detection algorithm [17]. This problem arises because certain algorithms, like modularity optimization, tend to merge smaller communities into larger ones, potentially overlooking meaningful structures at smaller scales.

According to El Ayeb et al. [18], evaluation metrics for overlapping community detection can be classified into intrinsic and extrinsic metrics. Intrinsic metrics evaluate structural properties of the identified communities, while extrinsic metrics compare the detected communities to a ground-truth. The most popular extrinsic metrics include the overlapping Normalized Mutual Information (ONMI), the Omega index, and the average F1-score.

The ONMI is an adaptation of the normalized mutual information for overlapping communities. It measures the similarity between two partitions, although it can be affected by the finite size effect, causing the average score to increase with the number of predicted communities.

The Omega index, adapted from the Adjusted Rand Index (ARI), considers both the observed and expected agreement between partitions. It is robust to the number of communities but has high computational complexity and performs poorly with multi-resolution partitions.

The average F1-score evaluates the harmonic mean of precision and recall for community pairs, providing a balanced measure of accuracy. However, it gives equal importance to precision and recall, and averaging can lead to high standard deviation.

El Ayeb et al. propose four new metrics to address limitations of existing methods: the inclusion rate, coverage rate, overlapping rate, and distribution rate. These metrics offer a comprehensive assessment of overlapping communities by considering the structure and similarity to ground-truth communities. The inclusion rate measures the embeddedness of result communities within ground-truth communities, the coverage rate assesses how well ground-truth communities are represented in the results, the overlapping rate evaluates the number of common nodes between communities, and the distribution rate compares the number of communities each node belongs to in the results and ground-truth.

Additionally, the Overlap Coverage metric, as defined by Ahn et al. [12], indicates how many densely overlapping communities were extracted by the algorithm. It counts the average number of memberships of the nodes who belong to communities. Overlap Coverage OC is defined as:

$$OC = \frac{\langle \sum ms(m_i) \rangle}{|S|} \tag{13}$$

where $m_i \in M = \bigcup c_i \subseteq C$ and C is the family of detected communities of nodes, such that $|c_i| > 2$, and ms is a function that counts the number of memberships of m.

By incorporating these metrics, a more accurate and meaningful evaluation of overlapping community detection algorithms can be achieved, providing deeper insights into the structure and quality of the detected communities.

1.3 Hypothesis

The central hypothesis of this research posits that pluralistic homophily in networks with overlapping communities can be accurately quantified through specific metrics that consider both global and local assortativity. Additionally, it is hypothesized that these metrics, within a comprehensive analytical framework, will provide a deeper understanding of connection patterns and the underlying dynamics in networks such as social, collaborative, and biological networks.

Building on this foundation, this thesis proposes a holistic analytical framework that includes two new metrics to quantify pluralistic homophily. The first metric assesses pluralistic homophily at the network level, based on the Pearson correlation coefficient applied to community memberships. The second metric evaluates pluralistic homophily at the local level, considering the behavior of individual nodes. This framework is further supported by traditional network analysis metrics, enhancing its ability to analyze pluralistic homophily comprehensively. These metrics were validated through an empirical analysis of nine diverse networks, revealing distinct patterns of pluralistic homophily and highlighting the need for a more nuanced understanding of node interactions in overlapping communities.

1.4 Scope of the Research

This thesis aims to develop a comprehensive analytical framework for studying pluralistic homophily, with the goal of validating the proposed metrics through empirical analysis of diverse networks, including social, collaborative, and biological networks. The research specifically focuses on undirected and unweighted networks, chosen for their prevalence in many practical applications and the clarity they provide in analyzing basic network properties. By doing so, the research conducted by this thesis seeks to clarify the nature of pluralistic homophily and provide a foundation for various applications in network science, such as link prediction, node classification, and recommendation systems.

While the scope is limited to undirected and unweighted networks, the analytical framework and methodologies developed lay a solid foundation for future research

on directed and weighted networks. The principles and metrics proposed here can be adapted and extended to handle the additional complexities of such networks, thereby broadening the applicability of pluralistic homophily analysis.

The framework integrates both novel and traditional network analysis metrics to offer a robust methodology for understanding complex interactions within overlapping communities, thereby enhancing the analytical capabilities and providing insights for future applications in the field of network science.

1.5 Organization of the Thesis

The thesis is organized as follows:

- **Abstract** - This section provides a summary of the thesis in Spanish and English, outlining the key findings and contributions.
- **Introduction** - This section introduces the research topic, outlines the problem statement, includes the theoretical background, presents the scope of the research, and states the hypothesis.
- **Objectives** - This section details the specific goals and aims of the research.
- **Related Works** - This section reviews the existing literature and research relevant to pluralistic homophily, including studies on assortativity, community detection, and overlapping communities. It provides context and background for understanding the contributions of this thesis.
- **Pluralistic Homophily** - This section presents the approach to measure pluralistic homophily and the framework for its analysis. The section covers two main contributions: the proposed approach for measurement and the development of a comprehensive analytical framework.
- **Experimental Setup** - This section describes the datasets used, the methodologies applied for the analysis, and the pipeline designed to conduct the related experiments.
- **Results and Discussion** - This section presents the findings from the experimental analysis of the nine datasets, including social, collaborative, and biological networks. It also provides an in-depth analysis of the results, discussing their implications and significance in the context of pluralistic homophily.
- **Conclusions and Future Work** - This section concludes the thesis with a summary of the findings and suggests directions for future research.

Additionally, the thesis includes the following appendices:

- **Local Pluralistic Homophily for the Text Network** - This appendix provides detailed analysis and findings of local pluralistic homophily specific to the text network dataset, which is used in the thesis as a didactic example to illustrate the proposed methods.
- **Residual Analysis for Linear and Quadratic Models** - This appendix presents a table with the residual values for the linear and quadratic models used in the study. These values are analyzed in the main body of the thesis to examine the presence of outliers and their impact on performance metrics such as AUC curves, supporting the robustness of the logistic regression model used.
- **Detailed Coefficients of Logistic Regression Models Across Networks** - This appendix lists the detailed coefficients of the logistic regression models applied across different networks, which are used to categorize local pluralistic homophily according to network centrality measures.
- **Preliminary Work: Overlapping Community Detection in StackOverflow** - This appendix includes preliminary work on overlapping community detection, specifically focusing on the StackOverflow dataset. It provides an initial exploration of pluralistic homophily metrics within this context.

2 Objectives

2.1 General Objective

The general objective of this work is to develop and validate a set of metrics to quantify pluralistic homophily in networks with overlapping communities, and to establish an analytical framework that integrates these metrics with other network metrics, providing a deeper understanding of community structure and node interactions.

2.2 Specific Objectives

- To introduce two metrics for measuring pluralistic homophily at both the network and node levels.
- To define an analytical framework that characterizes the relationship between pluralistic homophily and the network's community structure, as well as the roles and positions of nodes within the network.
- To apply the proposed metrics to diverse datasets, including social, collaborative, and biological networks, to demonstrate their generalizability and robustness.
- To validate the proposed metrics within the analytical framework by determining the correlation between pluralistic homophily and both network and community structures.
- To explore potential implications of pluralistic homophily for node attribute prediction, link prediction, and recommendation systems as future work.

3 Related Works

The concept of homophily has been a central theme in social network analysis, with assortativity measures commonly used to quantify the similarity between connected nodes [1]. However, traditional assortativity metrics do not account for the complexity of pluralistic interactions in networks with overlapping communities. Pluralistic interactions refer to the tendency of nodes to connect with others that share multiple community memberships, highlighting the multifaceted nature of their connections.

To address this gap, Yang and Leskovec [4] introduced the concept of pluralistic homophily, highlighting the tendency of nodes to connect with others who share mutual community memberships. Their approach revolutionizes traditional notions of similarity within networks by revealing how ground-truth communities overlap with each other. These findings uncover the phenomenon of pluralistic homophily, particularly evident within core-periphery structures, where nodes of high density frequently share multiple memberships with their neighbors. These nodes, which belong to the same communities as their neighbors, exhibit a high degree of similarity under one or more common attributes. This thesis builds on this foundational work by not only measuring pluralistic homophily from a holistic perspective but also developing an analytical framework to quantify its impact at both network and local levels.

In their experiments with ground-truth communities, Yang and Leskovec demonstrated the presence of pluralistic homophily across various datasets. This phenomenon is not only prevalent in social networks but also in biological networks, as they analyze protein interaction networks. This broad applicability underscores the importance of considering pluralistic homophily in diverse types of networks. In this thesis, we further explore this applicability by extending the analysis to a wider range of networks, including collaborative and informational networks, thereby validating and expanding the utility of pluralistic homophily in different contexts.

Yang and Leskovec also explain that traditional homophily operates in "pockets", suggesting that nodes with neighbors in different communities are less likely to share attributes. In contrast, this thesis demonstrates that pluralistic homophily reveals the similarity of nodes is proportional to the number of shared memberships or functions, rather than just their similarity along a single dimension. While Yang and Leskovec show a multidimensional perspective, revealing that the most central nodes in a network are those with the most shared community memberships, this thesis incorporates this perspective by examining how pluralistic homophily correlates with various centrality metrics, not just centrality and degree, thus providing deeper insights into the structural dynamics of networks.

Furthermore, their work delves into the global structure of core-periphery for-

mations. They illustrate how nodes in the core are highly interconnected and share many community memberships, while peripheral nodes have fewer connections and memberships. This global perspective provides a deeper understanding of how pluralistic homophily shapes the overall network structure. In relation to this, this thesis expands on these findings by analyzing how changes in community structure affect pluralistic homophily and how these changes can be used to predict network behaviors.

In addition to the foundational work by Yang and Leskovec, several other studies have mentioned or applied the concept of pluralistic homophily in various contexts. For example, [19] explores pluralistic homophily in different networks, analyzing the distribution of local pluralistic homophily and its relation to various structural and topological characteristics of a network. The study identifies significant patterns that enhance the understanding of how pluralistic homophily affects communities and suggests possible applications of local pluralistic homophily in future research. This previous work by the author of this thesis lays the groundwork and provides continuity to the current research, which aims to develop a comprehensive analytical framework for studying pluralistic homophily and validating its metrics through experimental analysis.

On the other hand, the integration of pluralistic homophily and node roles has been explored in recent studies to provide a comprehensive understanding of network structures. For instance, Costa and Ortale [20] developed two Bayesian probabilistic generative models, SCANNER and PERISCOPE, which incorporate the concept of pluralistic homophily along with node attributes and behavioral role patterns. These models aim to unify community detection and role analysis by leveraging node attributes and roles to explain link formation in networks. While Costa and Ortale in their work, focuses on integrating pluralistic homophily with node roles through probabilistic models using latent variables to capture node affiliations, this thesis explicitly quantifies pluralistic homophily with newly proposed metrics. This approach allows for a more granular analysis of how pluralistic homophily influences node interactions and network structure at both the network and local levels. Furthermore, the analytical framework developed in this thesis validates the importance of pluralistic homophily in network behaviors and extends its applicability to various real-world datasets, similar to the validation performed by Costa and Ortale.

Finally, Saha et al. [21] examined the role of overlapping groups in information dissemination within social networks. They modeled intergroup networks as random threshold graphs, where the presence of common members between groups significantly influences the spread of information. This aligns with the concept of pluralistic homophily, where overlapping community memberships play a crucial role in the connectivity and information flow within the network. Their analysis

of structural properties such as degree distribution, largest component size, edge density, and local clustering coefficient in intergroup networks further supports the understanding of how pluralistic homophily shapes network dynamics. This work emphasizes the importance of considering overlapping community structures in the study of network behaviors and information propagation. In relation to this thesis, it also analyzes the behavior of community structures, where Saha et al. focus on the dissemination of information within these community structures that exhibit pluralistic homophily. In contrast, this thesis explores the relationship between the phenomenon of pluralistic homophily and community structure through a framework that introduces centrality metrics. This approach provides a broader perspective on how behaviors such as information dissemination, among others, can be related to pluralistic homophily, offering deeper insights into the structural dynamics of networks.

4 Proposed Metrics and Framework for Measuring Pluralistic Homophily

Building on the foundational concepts of how nodes tend to connect based on their attributes, this thesis presents an approach to measure pluralistic homophily. While previous authors defined pluralistic homophily based on the number of shared community memberships, our approach focuses on the total number of community memberships of each node, providing a more holistic view of the phenomenon. Specifically, the number of memberships is quantified by counting the number of communities to which each node is affiliated within the network.

This thesis introduces two novel metrics (h and h_v). The first metric, h, applies the traditional correlation coefficient, represented by Equation 2, by incorporating the number of community memberships as a scalar value to determine the correlation between nodes at the network level. The second metric, h_v, evaluates pluralistic homophily at the local level, focusing on individual nodes and their connections, using Equation 12.

These metrics are complemented by well-established metrics in network theory, such as link density, overlap coverage, and centrality metrics, to provide a comprehensive framework for analyzing pluralistic homophily in various network contexts. The incorporation of these metrics in the analytical framework allows for a deeper understanding of how community structures and node centrality influence pluralistic homophily. At the network level, metrics like link density and overlap coverage help evaluate the overall coherence and strength of community structures, providing context for interpreting the h values. At the local level, centrality metrics offer insights into the relationship between the role and position of the nodes and h_v, enriching the analysis of local pluralistic homophily behaviors.

4.1 Network Pluralistic Homophily

Understanding pluralistic homophily at the network level requires an effective metric that can capture the tendency of nodes to connect with others based on their memberships. The traditional assortativity measure is adapted here to focus on how nodes with a similar number of memberships tend to link to each other. Rather than measuring the tendency of nodes to link with others with similar community memberships, this adaptation provides a holistic view of network interactions by instead considering the total number of communities to which each node belongs, treated as a scalar value.

This scalar value is then used in Equation 2 to quantify pluralistic homophily. The proposed measure of pluralistic homophily applies this equation by considering

the number of community memberships of nodes. The resulting equation is:

$$h = \frac{1}{\sigma_q^2} \sum_{xy} m_x m_y (e_{m_x m_y} - q_{m_x} q_{m_y}) \tag{14}$$

In this equation, m_x and m_y represent the number of community memberships of nodes x and y in a link, $e_{m_x m_y}$ is the joint probability for the membership values m_x and m_y, and q represents the membership distribution across the network. The term σ_q^2 is the variance of the membership distribution q, reflecting the variability in the number of memberships across nodes. Thus, while the original equation uses σ_a and σ_b as standard deviations, this version consolidates these into σ_q^2, the variance, for simplicity and relevance to the specific context of community memberships.

This application allows the measure to capture the broader range of homophilous behaviors based on the number of community memberships, providing a more holistic understanding of node interactions within the network.

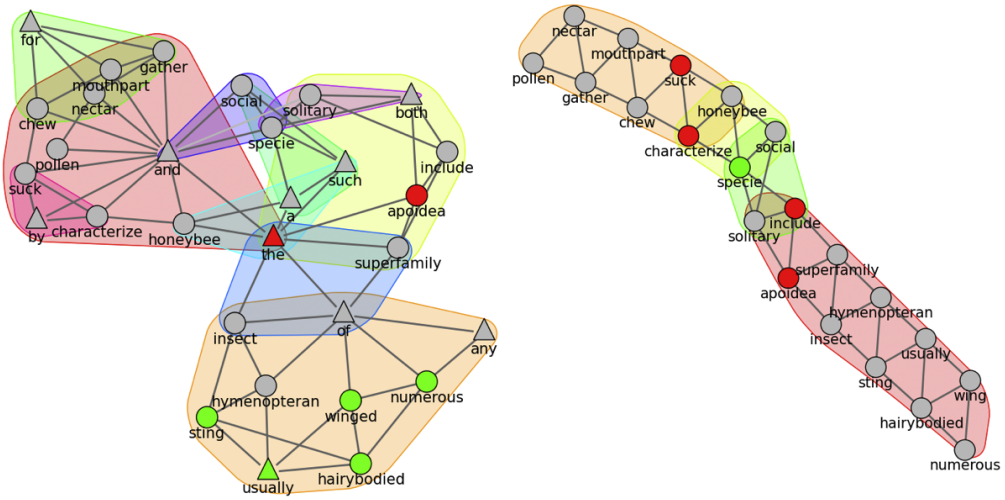
4.2 Local Pluralistic Homophily

The concept of local assortativity, introduced by [8], can also be applied to the field of pluralistic homophily. Local pluralistic homophily measures the tendency of individual nodes to connect with others that have a similar number of community memberships. This application allows for a detailed analysis of how individual nodes may exhibit homophilic tendencies that differ from the overall network tendencies. By focusing on the number of community memberships, local pluralistic homophily provides a more granular understanding of node interactions and their roles within overlapping community structures.

The metric of local pluralistic homophily extends from the concept of local assortativity, represented by Equation 12, using the number of community memberships as a scalar value. Specifically, the number of memberships is quantified by counting the number of communities to which each node is affiliated within the network. The scalar value for memberships is then used to define the local pluralistic homophily h_v of a node v as follows:

$$h_v = \frac{M^{-1}}{2\sigma_q^2} \left[\sum_{i=1}^{d_v} (m_v - \mu_q)(m_i - \mu_q) \right] \tag{15}$$

In this new equation, i loops through all neighbors of node v, m_v is the number of memberships of node v, and m_i is the number of communities to which neighbor node i belongs. M is the total number of nodes in the network, and q is the probability distribution of m. Consequently, h_v represents the tendency of a specific node v to link to other nodes with a similar number of memberships. The sum of the h_v values



a) $h = 0.2057$ *b*) $h = 0.6388$

Figure 2: Network visualization of word adjacency from a text defining a bee, highlighting community structures. (a) The network with all words, showing a moderate level of pluralistic homophily ($h = 0.2057$), with overlapping communities identified by colored shading. In this network, nodes whose words are identified as stopwords appear in a triangle shape. (b) The network with the stopwords removed, revealing a higher degree of pluralistic homophily ($h = 0.6388$). In both networks, nodes are color-coded to reflect their pluralistic homophily classification: green for assortative, gray for non-assortative, and red for disassortative nodes.

for all nodes is consistent with the results obtained using Equation 14 for pluralistic homophily of the entire network.

This formulation ensures that local pluralistic homophily captures the homophilic behaviors based on the number of community memberships, providing a nuanced understanding of node interactions within the network.

4.3 An Illustrative Example of Pluralistic Homophily

Figure 2 presents the visualizations of two networks built from a text that defines a bee in a school dictionary. In the network, nodes are words and the links are established based on the adjacency of words within a window of size k.

The networks highlight overlapping communities identified by colored shadows, with Figure 2a representing the network generated from the full text and Figure 2b derived from the text with stopwords removed, processed using the NLTK library [22]. The network formed with all words demonstrates assortative mixing by node memberships, that is, pluralistic homophily ($h = 0.2057$), while the network without stopwords exhibits higher pluralistic homophily ($h = 0.6388$). In both networks, nodes are color-coded: green for nodes with assortative local pluralistic homophily,

gray for non-assortative, and red for disassortative. The pluralistic homophily of nodes is categorized using the mean (μ) and standard deviation (σ) values. Nodes with h_v values above $\mu + \sigma$ are considered assortative in terms of their pluralistic homophily, while those below $\mu - \sigma$ are disassortative. Nodes with h_v values within these limits are non-assortative. Appendix A presents the local pluralistic homophily values for all nodes in both networks.

Notably, some nodes exhibit assortative and non-assortative local pluralistic homophily, contrary to the overall pluralistic homophily of the network. In both networks, most nodes are non-assortative, highlighting the importance of analyzing pluralistic homophily at various levels. Another significant observation concerns the impact of stopwords on the structure of both networks and their respective community structures. Stopwords, characterized by their high frequency of use across various texts, can exhibit a high degree within the resulting word networks if not removed. In the network containing stopwords, these words are situated in different areas of community overlap, showcasing diverse values of pluralistic homophily. This ranges from assortative (e.g., 'usually'), through non-assortative (e.g., 'a', 'and', 'any', 'both', 'by', 'for', 'of', 'such'), to disassortative ('the'). Additionally, stopwords display varying levels of community overlap, highlighting the complexity of their roles within the network structure. This behavior of the nodes, with varying levels of overlap, positions within the network, and pluralistic homophily, motivates the detailed exploration in this work. The aim is to dive deeper into this phenomenon, seeking to understand how these characteristics influence the formation and structure of communities in complex networks.

4.4 Categorization of Local Pluralistic Homophily

In the previous example, the mean of h_v and its standard deviation were used to establish the upper and lower limits of a range in which the values of local pluralistic homophily can be considered non-assortative. Consequently, values above and below these limits are considered assortative and disassortative, respectively. However, this criterion for categorizing the pluralistic homophily of nodes can be subjective. Previous studies have explored the relationship between assortativity and network size, showing that assortativity can be influenced by network size, converging to zero in large networks [23] [24], which affects its mean and standard deviation in the same manner. In contrast, in a small network, the standard deviation may be relatively large due to the smaller amount of data. Since the calculation of local pluralistic homophily is based on network assortativity—in this case, assortativity based on the number of memberships of a node—and as seen in Equation 15, local pluralistic homophily includes the term M, which defines the number of nodes in the

network. It is therefore understandable that network size impacts the calculation of pluralistic homophily values. Consequently, an adaptive approach to categorizing local pluralistic homophily is necessary to ensure meaningful classification across networks of different sizes. To achieve this, a factor k is defined and applied to the standard deviation σ to determine the limits for categorizing h_v. To account for different network sizes, three adaptive constants are introduced: k_s for small networks, k_m for medium networks, and k_l for large networks. The general formula for categorizing h_v using these adaptive constants is:

- Assortative: $h_v > \mu + k_x \cdot \sigma$
- Non-Assortative: $\mu - k_x \cdot \sigma \leq h_v \leq \mu + k_x \cdot \sigma$
- Disassortative: $h_v < \mu - k_x \cdot \sigma$

where k_x represents the adaptive constant corresponding to the network size (small, medium, or large). This adaptive approach ensures that the categorization limits are suitable for the specific characteristics of each type of network, providing a more precise and balanced view of pluralistic homophily.

4.5 Particular Cases of Pluralistic Homophily: Measuring with shared communities

Building on the work of Yang and Leskovec in [4], which explored the tendency of nodes to connect based on shared community memberships, a potential extension to the current metrics of pluralistic homophily is outlined. While the metrics h and h_v capture the tendency of nodes to connect with others having a similar number of community memberships, new metrics h' and h'_v are proposed to measure the specific overlap in community memberships between connected nodes.

The metrics h' at the network level can be defined as follows:

$$h' = \frac{1}{\sigma_q^2} \sum_{xy} \tilde{m}_{xy} (e_{\tilde{m}_{xy}} - q_{\tilde{m}_x} q_{\tilde{m}_y}) \tag{16}$$

where $\tilde{m}_{xy}$ is the number of shared community memberships for nodes x and y. All other terms are consistent with those defined for h in Equation 14.

Similarly, the metric h'_v at the local node level can be defined as follows:

$$h'_v = \frac{M^{-1}}{2\sigma_q^2} \left[\sum_{i \in N_v^*} (\tilde{m}_v - \mu_q)(\tilde{m}_i - \mu_q) \right] \tag{17}$$

where $\tilde{m}_v$ is the number of memberships that the node v shares with its neighbors, and $\tilde{m}_i$ is the number of communities that the neighbor i shares in turn with its

own neighbors. N_v^* denotes the set of neighbors of node v that share at least one community membership with v. All other terms are consistent with those defined for h_v in Equation 15.

Considering shared community memberships, h' and h'_v provide direct measures of the tendency of nodes to connect with others based on actual community overlap, rather than merely the number of memberships. These metrics align more closely with the phenomenon of pluralistic homophily as originally described by Yang and Leskovec, where nodes with shared community memberships are more likely to connect.

An important aspect of these metrics is that h and h_v will always be greater than or equal to h' and h'_v, respectively. This is because h and h_v consider all community memberships, while h' and h'_v focus solely on shared memberships. This relationship underscores the generality of the metrics h and h_v, demonstrating their ability to encompass a broader range of homophilic behaviors and to particularize into more specific cases.

While h' and h'_v are not explored in this study, future work will investigate their application and effectiveness in capturing the specific homophilic tendencies observed in different network contexts.

4.5.1 Proof that h' is a particular case of h

The relationship between two metrics of pluralistic homophily, h and h', is explored. The goal is to demonstrate that the metric h' is a finer measure than h. To achieve this, definitions for the concepts used are provided, the lemma is stated, and a proof by contradiction is presented.

Definitions. For clarity in the demonstration, some definitions are repeated here. Consider a finite set of nodes V_G ($|V_G| < \infty$) in a graph G. A *community* is a subset of nodes, denoted as $c \subseteq V_G$, where V_G is the set of all nodes in the graph G. The detected communities in a graph form a set $C = \{c_1, c_2, \ldots, c_n\}$ such that each c_i is a subset of V_G (i.e., a family of sets where each set is in turn a set of vertices of V_G).

The metrics of pluralistic homophily h and h' are defined as follows:

$$h = \frac{1}{\sigma_q^2} \sum_{x,y \in V_G} m_x m_y \left(e_{m_x m_y} - q_{m_x} q_{m_y} \right)$$

where m_x and m_y are the number of memberships of nodes x and y, respectively, and $e_{m_x m_y}$ is a measure of the expected interaction between nodes x and y based on their memberships, while q_{m_x} and q_{m_y} are the probabilities associated with these memberships.

$$h' = \frac{1}{\sigma_q^2} \sum_{x,y \in V_G} \tilde{m}_{xy} \left(e_{\tilde{m}_{xy}} - q_{\tilde{m}_x} q_{\tilde{m}_y} \right)$$

where $\tilde{m}_{xy}$ is the number of shared memberships of nodes x and y, and $e_{\tilde{m}_{xy}}$ is a measure of the expected interaction based on these shared memberships.

Lemma. For any set of communities C, the values of $\tilde{m}_{xy}$ in the equation for h' are always less than or equal to the values of $m_x m_y$ in the equation for h. Formally,

$$\tilde{m}_{xy} \leq m_x m_y \quad \forall\, x, y \in V_G \text{ that belong to at least one community } c \in C.$$

Proof by Contradiction. To prove that h' is finer than h, a proof by contradiction is used. Assume the opposite, that h' is not finer than h. This would imply that there exists at least one set of communities C where there is a pair of nodes x and y such that $\tilde{m}_{xy} > m_x m_y$.

However, by the lemma, $\tilde{m}_{xy} \leq m_x m_y$ since the number of shared memberships cannot exceed the product of the individual memberships of x and y. Moreover, since $e_{m_x m_y}$ and $e_{\tilde{m}_{xy}}$ are measures of the expected interaction based on these membership values, the relationship holds:

$$\tilde{m}_{xy} \left(e_{\tilde{m}_{xy}} - q_{\tilde{m}_x} q_{\tilde{m}_y} \right) \leq m_x m_y \left(e_{m_x m_y} - q_{m_x} q_{m_y} \right).$$

This contradicts the initial assumption that $\tilde{m}_{xy} > m_x m_y$. Therefore, the assumption that h' is not finer than h leads to a contradiction with the definition of $\tilde{m}_{xy}$. Thus, the initial assumption is false, and consequently, h' is indeed finer than h.

4.6 Framework for Pluralistic Homophily Analysis

Understanding the relationship between pluralistic homophily and community structures is crucial for comprehending the dynamics of complex networks.

As shown previously in the example of the network for the text bee definition, pluralistic homophily trends at the network level do not necessarily reflect the patterns observed at the node level. This discrepancy suggests that pluralistic homophily is influenced by the network and community structure, as well as the position of individual nodes within the network. This gap necessitates the development of an analytical framework capable of integrating these interactions and providing a deeper understanding of node connectivity.

This thesis proposes an analytical framework that systematically integrates pluralistic homophily metrics with other key metrics in the analysis of complex networks to achieve a comprehensive understanding of the pluralistic homophily phenomena.

Initially, the framework explores the influence of community overlap structures and their impact on pluralistic homophily, focusing on how it is affected when the overlapping has extensive coverage or not, which could result in many nodes with a high or lower number of memberships, respectively. Subsequently, the relationship between the pluralistic homophily of a node and its position and role within the overall network structure is examined.

By incorporating both network-level and local-level analyses, the framework seeks to uncover hidden patterns and relationships that traditional methods overlook. This comprehensive approach provides a deeper understanding of pluralistic homophily within networks, helping to interpret community structures within the broader context of complex networks.

4.6.1 Network-Level Analysis

The measurement of pluralistic homophily is influenced by two key factors: the number of community memberships per node and the connections between these nodes (see Equation 14). This underlines the importance of accurately identifying communities within networks, either through ground-truth communities or through algorithms designed to detect overlapping communities. Understanding the influence of community structures on network behaviors, particularly in networks with overlapping communities, has been a significant focus in network science. Community detection methods have evolved to uncover unique structural insights, ranging from clique-based approaches to link clustering and label propagation [25].

Prominent algorithms such as CFinder [26], Bigclam [27], and HLC [12] stand out, particularly for their effectiveness in uncovering overlapping community structures. Given that each algorithm detects different sets of communities, it is necessary to consider the performance of each in terms of the resulting community structure. Several methods have been developed for community detection, such as clique-based approaches, link clustering, and label propagation, each designed to reveal unique structural insights [25]. Prominent algorithms such as CFinder [26], Bigclam [27], and HLC [12] have been particularly effective in revealing these community structures.

Metrics to assess the quality of detected communities, such as normalized mutual information (NMI) [28], the extended normalized mutual information for overlapping community detection (ONMI) [29], and the Omega Index [30], have been proposed by various authors [18, 31]. Despite the importance of such metrics, this research takes a different approach by focusing on both the overlap between communities

and the intrinsic community structure, whether resulting from community detection algorithms or established as ground-truth. Therefore, metrics like Overlap Coverage (OC) and Scaled Link Density ($\tilde{\rho}$) are used, which measure the resulting structure and not the quality of the detected communities per se.

To address the complexity of overlapping communities, metrics such as Overlap Coverage (OC) [12] and Scaled Link Density ($\tilde{\rho}$) [32] have been developed. Although OC has been introduced previously in this thesis, it is revisited here for clarity. Overlap Coverage (OC) assesses the extent to which nodes belong to multiple communities, reflecting the true level of overlapping within the communities. Formally, OC is defined as:

$$OC = \frac{\sum_{i=1}^{N} M_i}{N} \tag{18}$$

where M_i is the number of communities to which node i belongs, and N is the total number of nodes in the network.

Scaled Link Density ($\tilde{\rho}$) measures the internal connectivity strength within communities. This metric is defined for a single community as:

$$\tilde{\rho} = \frac{2t}{s(s-1)} \tag{19}$$

where t represents the number of internal links within the community and s is the number of nodes in that community.

By integrating the OC and $\tilde{\rho}$ metrics into the analytical framework, the two main factors influencing pluralistic homophily are addressed. The OC allows for analysis of its variation with respect to the level of community overlap (as defined in 18), while $\tilde{\rho}$ reveals the strength of connectivity between nodes within the same community (as defined in 19). Together, these metrics enhance the understanding of how nodes with similar community memberships tend to connect, offering deeper insight into the dynamics of pluralistic homophily.

4.6.2 Local-Level Analysis

Nodes with high pluralistic homophily can occupy both core and peripheral positions within the network, indicating that pluralistic homophily influences not only the connections of nodes but also their strategic positioning within the overall network structure. Exploring how these nodes' placements relate to their roles and potential influence within the network is crucial.

Research by [33] analyzes the proximity between the hubs and the overlapping nodes within networks, suggesting a significant interaction between the centrality of the nodes and the structures of the community. Their findings indicate that hubs

are often located near nodes that are central to multiple communities, revealing a complex interplay between nodal centrality and community architecture. Another study by [4] characterizes overlapping regions within community structures, suggesting that nodes in overlapping areas between communities are more densely connected than those in non-overlapping areas. This highlights that community hubs, or nodes with significant connections within one or multiple communities, are often located within overlapping areas, with the probability of a hub being in an overlap area increasing with the size of the overlap.

Moreover, research by [34] further supports this approach by introducing a generative model that combines preferential and anti-preferential attachment to create networks with realistic assortativity and hierarchical clustering. This model illustrates how nodes that initially connect to low-degree nodes can become high-degree hubs, providing insights into the structural evolution of social networks. Such findings are relevant to the study of pluralistic homophily, where the number of community memberships plays a crucial role in node connectivity.

Building on these insights, it becomes evident that centrality measures are essential for understanding the roles and influence of nodes within networks, particularly in the context of pluralistic homophily, where the position and influence of a node can significantly affect its homophilic behaviors. The analytical framework integrates then various centrality metrics to quantify the strength and nature of these interactions. These centrality metrics are:

Degree centrality (C_d) is defined as:

$$C_d(v) = \frac{deg(v)}{N-1} \tag{20}$$

where $deg(v)$ is the number of direct connections that a node v has, and N is the total number of nodes. High degree centrality indicates nodes with many direct connections, often acting as hubs within their communities. Analyzing degree centrality helps in understanding whether nodes with high pluralistic homophily are also highly connected hubs or if they occupy more peripheral positions.

Closeness centrality (C_c) is given by:

$$C_c(v) = \frac{1}{\sum_{u \neq v} d(v,u)} \tag{21}$$

where $d(v,u)$ represents the shortest path distance between nodes v and u. This metric assesses how central a node is within the network by measuring the average shortest path length from the node to all other nodes. Nodes with high closeness centrality are well-positioned to spread information quickly across the network. This metric helps determine if nodes with high pluralistic homophily are centrally located within the network or if they are more isolated.

Betweenness centrality (C_b) is defined as:

$$C_b(v) = \sum_{s,t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{22}$$

where σ_{st} represents the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ those passing through v. This metric quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Nodes with high betweenness centrality often play critical roles in connecting different parts of the network. This metric is crucial for identifying whether nodes with high pluralistic homophily also serve as key intermediaries within the network.

Eigenvector centrality (C_e) is calculated by:

$$C_e(v) = \frac{1}{\lambda} \sum_{t \in M(v)} a_{vt} C_e(t) \tag{23}$$

where $M(v)$ are the neighbors of v, a_{vt} is the adjacency matrix entry, and λ is a constant. This measure considers both the quantity and quality of a node's connections by accounting for the centrality of its neighbors. Nodes with high eigenvector centrality are influential not only due to their direct connections but also because they are connected to other highly central nodes. This measure helps analyze if nodes with high pluralistic homophily are also the most influential within the network.

To quantify the impact of these centrality measures on nodes' pluralistic homophily, a logistic regression model is implemented. This model provides a robust framework to predict how these traits influence nodes' likelihood of forming connections within similar community structures. Each centrality metric served as an independent variable in a separate logistic regression model, with the pluralistic homophily category as the dependent variable. Using centrality measures effectively predicts the level of pluralistic homophily based on the structural attributes of the nodes. This comprehensive approach ensures a deeper understanding of the interplay between nodal centrality and pluralistic homophily at the local level. The logistic regression model, incorporating network structure, follows the principles outlined by [35], demonstrating how network-based logistic regression can improve classification accuracy by leveraging additional network information.

5 Experimental Setup

In this section, the methodological approach taken to explore the complex interactions between pluralistic homophily and community structures in various types of networks is described. The methodology focuses on conducting a detailed and systematic analysis to examine the variation of pluralistic homophily from the network level down to the dynamics at the level of individual nodes.

A set of metrics specifically designed to quantify pluralistic homophily is introduced. These metrics measure pluralistic homophily both at the network level (h) and at the local level (h_v). These metrics are integrated into a comprehensive analytical framework that incorporates other relevant metrics such as Overlap Coverage (OC) and local assortativity ($\tilde{\rho}$) at the network level, as well as centrality measures such as degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality at the local level, to provide a deeper understanding of community structures and node interactions. The framework facilitates a multi-level analysis, allowing for a detailed examination of how pluralistic homophily manifests across different network structures and the roles individual nodes play within these networks.

The following subsections elaborate on the specific metrics used, the analytical framework developed for this study, the datasets selected, and the pipeline implemented for the analysis.

5.1 Datasets of Selected Networks

Nine datasets of real-world networks and their respective communities were selected. These networks vary in size, structure, and community composition, providing a comprehensive understanding of pluralistic homophily in diverse contexts. Below, each network is described in terms of its connections and communities.

- **StackOverflow (SO)**: A collaborative question-and-answer network focused on programming and related topics. Nodes represent users, and edges represent the questions asked or answered between them. Communities reflect common interests, defined by tags associated with technology topics. Overlap occurs when a user participates in multiple communities due to their interest in various technologies.

- **DBLP**: A coauthorship network of scientific publications. Nodes represent authors, and edges indicate co-authorship. Communities are defined by publication venues such as journals and conferences. Overlap occurs when an author publishes in multiple venues.

- **Amazon**: A product co-purchasing network where nodes represent products and edges are formed when products are bought together. Product categories define communities; for example, books, electronics, and clothing each form distinct communities based on co-purchasing patterns. Overlap occurs when products from different categories are frequently bought together, indicating cross-category purchasing behaviors. Thus, a product can be classified into multiple categories, and the overlap is measured based on the number of categories shared between connected products.

- **LiveJournal**: An online blogging community. Nodes represent users, and edges represent explicit relationships established between them. Communities are formed by groups created by users that others can join. Overlap occurs when a user belongs to multiple groups.

- **YouTube**: A video-sharing network. Nodes represent users, and edges represent friendship relationships. Communities are formed by groups to which users subscribe, and overlap occurs when a user subscribes to multiple groups.

- **Orkut**: A social network. Nodes represent users, and edges represent explicit relationships established between them. Communities are formed by groups created by users that others can join. Overlap occurs when a user belongs to multiple groups.

- **PPI**: A network mapping interactions between proteins in a cell. Nodes represent proteins, and edges indicate physical interactions between them. Communities are defined by sets of functionally interacting proteins.

- **DDI**: A network showing interactions between different drugs. Nodes represent drugs, and edges indicate the interactions between them. Communities are formed by groups of drugs with significant interactions.

- **Celegans**: A network that represents biochemical reactions in the metabolism of the nematode *Caenorhabditis elegans*. Nodes represent metabolites, and edges represent metabolic reactions. Communities are defined by sets of interrelated metabolic reactions.

The basic network metrics for the selected datasets were calculated. For communities, ground-truth communities available in the SNAP [36] and BioSNAP [37] websites were used, except for SO, PPI, DDI, and C.elegans, where the HLC algorithm [12] was applied for community detection. Using ground-truth communities ensures that the results are not affected by the noise or perturbations that can arise from inaccuracies in community detection algorithms. This allows for a more

accurate and reliable analysis of the networks. For datasets without ground-truth communities, the application of the HLC algorithm provides a consistent method for detecting communities, although it may introduce some variability compared to datasets with predefined community structures.

5.2 Pipeline for Pluralistic Homophily Analysis

For empirical analysis, a comprehensive pipeline is implemented following a structured step-by-step approach to process data, detect communities, calculate various metrics, and analyze correlations. The detailed outline of this process is as follows:

1. **Dataset Reading and Network Reconstruction:** Datasets are read to reconstruct network graphs. For larger networks such as LiveJournal, YouTube, and Orkut, sampling techniques are applied to effectively manage execution times during subsequent calculations.

2. **Community Detection:**

   (2a) *Using the HLC Algorithm:* In networks like SO, PPI, DDI, and C.elegans, communities are detected using the HLC algorithm. Multiple community structures are generated by varying the cut-off threshold t in the dendrogram.

   (2b) *Using Ground-Truth Data:* In networks such as DBLP, Amazon, LiveJournal, YouTube, and Orkut, communities are identified based on the ground-truth data.

3. **Metric Calculation:**

   (3a) *Network Level:* Overall pluralistic homophily h, Overlap Coverage (OC) (Equation 18), and Scaled Link Density ($\tilde{\rho}$) (Equation 19) across communities are computed.

   (3b) *Node Level:* Metrics such as h_v and centrality measures C_d (Equation 20), C_c (Equation 21), C_b (Equation 22), and C_e (Equation 23) are calculated for individual nodes.

4. **Correlation Analysis and Visualization:** The linear and monotonic relationships between centrality metrics and pluralistic homophily at both the network-level are initially examined using the Spearman correlation index. At the local-level, logistic regression models are then utilized to predict the categories of pluralistic homophily (assortative, non-assortative, disassortative) based on the centrality measures calculated for each node.

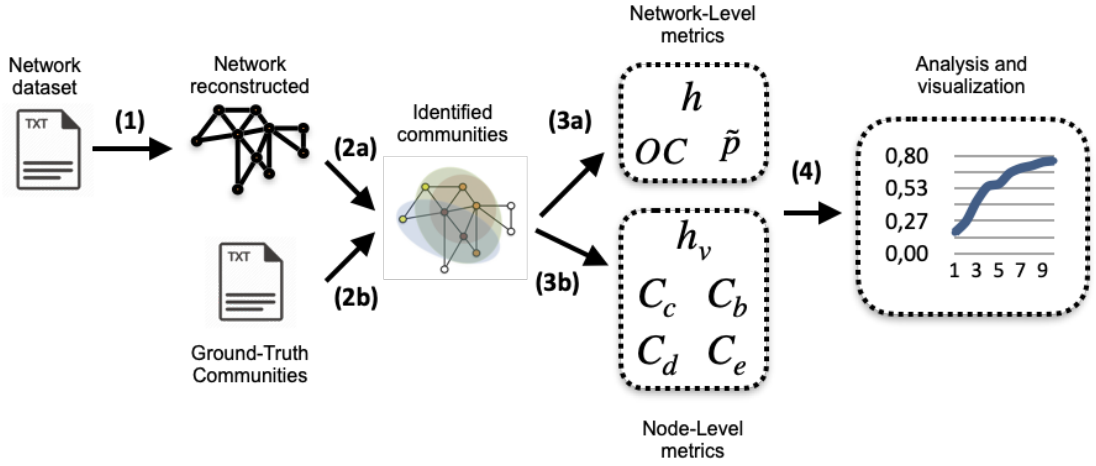


Figure 3: Experimental Pipeline. A sequential representation from dataset reading to correlation analysis and results visualization. The steps include network reconstruction (1), community detection using HLC (Step 2a) and ground-truth data (Step 2b), metric calculation at the network-level (Step 3a) and local-level (Step 3b), and correlation analysis with visualization (Step 4), reflecting the detailed process implemented in Python.

Figure 3 illustrates the pipeline from data reading to the visualization of the final results. This pipeline corresponds to the process described earlier, detailing each step from dataset reconstruction to the correlation analysis and visualization.

The entire framework analysis is coded in Python, utilizing various libraries for distinct purposes: `numpy`, `pandas`, `scikit-learn`, and `Matplotlib` provide computational support and visualization capabilities; `igraph` [38] is used for graph manipulation and analysis, `littleballoffur` [39] assists in sampling large networks, and `networkit` [40] supports several tasks on large networks. The complete suite of programs is available on GitHub at `https://github.com/fernandobarraza/pluralisticHomophily`.

5.3 Generating Multiple Community Structures

To analyze the relationship between overlapping communities and pluralistic homophily, the Hierarchical Link Clustering (HLC) algorithm will be used. By adjusting the cut-off threshold t in the dendrogram that represents the hierarchical clustering of network links, different sets of overlapping communities C_t will be generated, each with configurations of varying numbers of communities, degrees of overlap, and other structural characteristics. This approach provides the necessary variety in community configurations to thoroughly examine their impact on pluralistic homophily. Following this, for each community set, the pluralistic homophily of the network h (Equation 14) will be calculated, and the correlation of this metric with Overlap Coverage (OC) (Equation 18) and Scaled Link Density ($\tilde{\rho}$) of the

community set detected in the network will be examined.

Once the communities for each threshold are identified, the relationship between pluralistic homophily and the OC and $\tilde{\rho}$ measures will be evaluated using Spearman's correlation coefficient. This coefficient is a non-parametric measure that evaluates how the relationship between two variables can be described by a monotonic function, without assuming a linear relationship between them. A positive Spearman correlation between pluralistic homophily and OC would suggest that higher community overlap is associated with higher pluralistic homophily. Conversely, a negative correlation would indicate that as overlap increases, pluralistic homophily decreases. The Spearman coefficient is chosen because of its robustness against non-normal distributions and its lower sensitivity to outliers [41, 42].

Figure 4 demonstrates the application of varying thresholds to the dendrogram of the HLC algorithm, revealing the dynamic community configurations within a text network previously discussed. This analysis underscores the significant variations in homophily and overlap coverage among communities that can arise from even slight adjustments in the threshold parameter. A lower threshold results in a community structure with moderate link density and significant overlap coverage, whereas a higher threshold exposes a more diverse community membership characterized by looser connections between nodes. These findings highlight the intricate dynamics that influence the segmentation of the network into communities.

At a threshold $t = 0.3$, the network is divided into $|C_{t=0.3}| = 3$ communities with pluralistic homophily $h(t = 0.3) = 0.1130$, Overlap Coverage (OC) of $OC(t = 0.3) = 0.8113$, and average link density $\langle\tilde{\rho}(t = 0.3)\rangle = 0.5177$. At the optimal threshold t^*, $|C_{t^*}| = 10$ communities are identified, exhibiting a pluralistic homophily of $h(t^* = \max(\overline{p})) = 0.0382$, with $OC(t^*) = 0.9245$ and $\langle\tilde{\rho}(t^*)\rangle = 0.6658$. Increasing the threshold to $t = 0.5$ results in $|C_{t=0.5}| = 13$ communities, a pluralistic homophily $h(t = 0.5) = 0.0124$, Overlap Coverage $OC(t = 0.5) = 1.0$, and average link density $\langle\tilde{\rho}(t = 0.5)\rangle = 0.8608$. These configurations illustrate the flexibility of the network and the impact of threshold selection on community detection, showing how varying the threshold in HLC provides different community structures with distinct OC and $\tilde{\rho}$ measures. This variability enables a comprehensive analysis of pluralistic homophily behavior across diverse community structures.

5.4 Logistic Regression Model for Centrality Measures and Pluralistic Homophily

To understand the influence of nodal centrality on pluralistic homophily at a local level, a logistic regression model is implemented. This model takes four centrality measures—degree centrality (C_d), closeness centrality (C_c), betweenness centrality

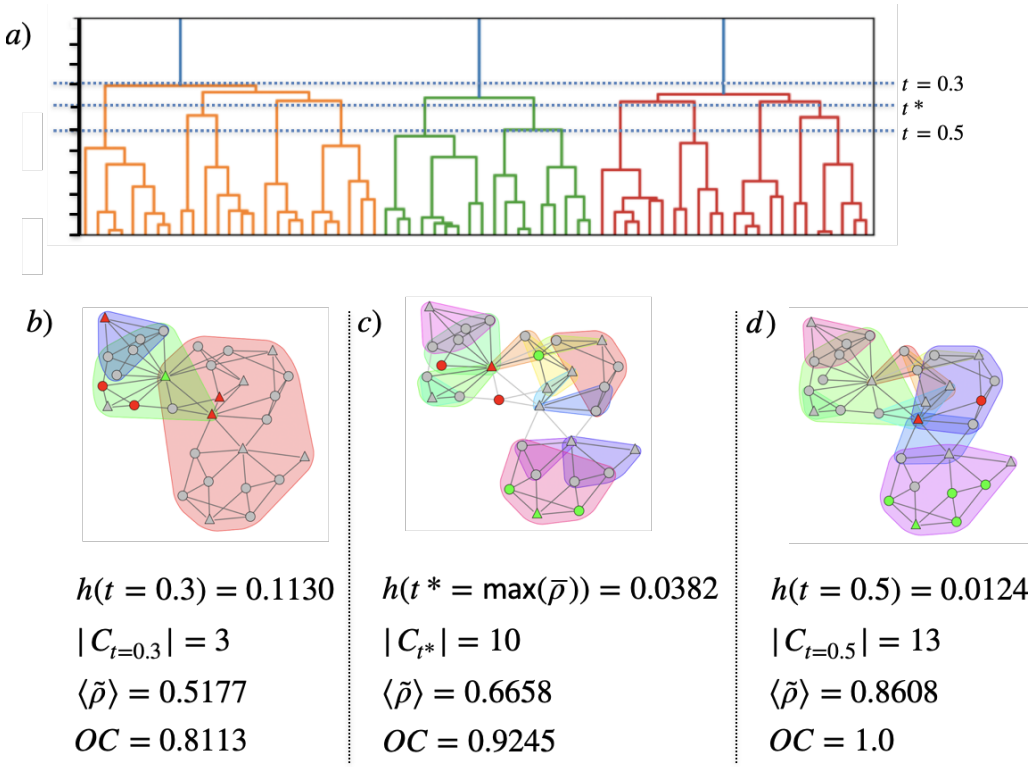


Figure 4: Visualization of Hierarchical Link Clustering (HLC) at different thresholds. (a) Dendrogram representation of HLC with cutoff lines at thresholds $t = 0.3$, t^*, and $t = 0.5$, which are used to determine community structures at various levels of granularity. (b) At $t = 0.3$. (c) At the optimal threshold t^*. (d) At $t = 0.5$.

(C_b), and eigenvector centrality (C_e)—as independent variables, and the pluralistic homophily category (h_v) as the dependent variable. In this context, the pluralistic homophily category is a categorical variable indicating whether a node is assortative, non-assortative, or disassortative.

The logistic regression model is designed to predict the likelihood of a node belonging to a specific homophily category based on its centrality measures. For instance, in the example of a text network derived from the definition of the bee, nodes represent words, and edges represent co-occurrences of words within a predefined context window. Communities are detected using the HLC algorithm, and each word (node) is assigned a pluralistic homophily category based on its connections within and across these communities.

The steps for implementing the logistic regression model are as follows:

1. **Feature Extraction:** Calculate the centrality measures (C_d, C_c, C_b, C_e) for each node in the network.

2. **Label Assignment:** Assign each node a pluralistic homophily category (h_v) based on its community memberships and interactions.

3. **Model Training:** Use the centrality measures as input features and the homophily categories as output labels to train the logistic regression model. Each centrality measure is included as an independent variable in the model.

4. **Prediction:** Apply the trained model to predict the homophily category for each node based on its centrality measures.

5. **Evaluation:** Assess the model's performance using metrics such as the ROC curve and AUC, which provide insights into the model's ability to distinguish between different homophily categories.

By integrating these steps, the logistic regression model provides a robust framework to analyze the interplay between nodal centrality and pluralistic homophily. This approach allows for a detailed examination of how different centrality measures influence the likelihood of a node exhibiting specific homophilic behaviors within the network.

5.4.1 Data Preparation

The preparation of data is a critical step in ensuring the robustness and accuracy of the logistic regression model. This involves identifying and utilizing appropriate community structures, sampling nodes, calculating centrality measures, and stratifying nodes based on their pluralistic homophily. The following outlines the detailed steps taken to prepare the data for analysis.

Where ground-truth community structures are available, they provide a benchmark for comparison; in their absence, communities identified by the Hierarchical Link Clustering (HLC) algorithm are used. This dual approach enriches the analysis, allowing the assessment of the robustness of findings in various types of community structures derived from different sources. Specifically, ground-truth communities are used for networks such as DBLP, Amazon, LiveJournal, YouTube, and Orkut, while HLC-detected communities are used for networks such as SO, PPI, DDI, and C.elegans [12]. The analysis does not enforce a specific cut threshold for the dendrogram but instead uses the cut that maximizes the average link density, denoted by t^*.

Within these network structures, nodes are sampled and categorized into assortative, non-assortative, and disassortative categories based on their local pluralistic homophily, according to the criteria defined in Section 4.4. Local pluralistic homophily (h_v) and a suite of centrality metrics—degree centrality (C_d), closeness centrality (C_c), and eigenvector centrality (C_e)—are computed. Notably, the computation of betweenness centrality (C_b) poses significant computational challenges, especially in large networks. An 'Approximated Betweenness' approach is employed,

using algorithms that significantly reduce the computational load of calculating betweenness centrality. Pioneering work has shown that such approaches can significantly reduce computational complexity, operating in $O(nm + n^2 \log n)$ time for both unweighted and weighted networks, where n is the number of nodes and m is the number of edges [43]. This method ensures both computational efficiency and reasonable approximation accuracy in the analyses [44].

The model is trained using a subset of data, specifically 70% of the nodes chosen at random, ensuring a balanced representation of each homophily category. The remaining 30% of the data are used to validate the models, assessing their accuracy and robustness through standard statistical metrics such as AUC (Area Under the Curve) for ROC (Receiver Operating Characteristic) analysis and confusion matrices. To ensure a fair comparison and improve the model's interpretability, all centrality metrics are normalized before inclusion in the regression models.

In the subsequent section, the results of this analysis, including the ROC-AUC curves for the text network example, will be presented to illustrate the effectiveness of the logistic regression model in predicting pluralistic homophily categories based on nodal centrality measures.

6 Results

This section presents the results of the empirical analysis conducted to understand the dynamics of pluralistic homophily within various network structures. The analysis is divided into several key parts to provide a comprehensive understanding of pluralistic homophily at both network and node levels.

First, an overview of how pluralistic homophily was measured is provided, detailing the methods and metrics used. The results then focus on network-level pluralistic homophily, analyzing community overlap and its impact on pluralistic homophily. This includes an examination of the relationship between community structures and pluralistic homophily, presenting key findings and illustrating how different community configurations affect homophilic tendencies.

Local-level pluralistic homophily results are then explored, discussing the Probability Density Function (PDF) of pluralistic homophily values across different networks to offer insights into the distribution of homophily values. Additionally, the Cumulative Distribution Function (CDF) analysis is examined to understand the cumulative behavior of pluralistic homophily in the networks studied.

Furthermore, the relationship between local pluralistic homophily and community memberships is investigated, as well as the correlation between local pluralistic homophily and various centrality metrics. Key insights and patterns observed from these analyses are highlighted, emphasizing the role of individual nodes within the broader network context.

This structured approach ensures a thorough analysis of pluralistic homophily, offering a detailed understanding of its role and significance in complex networks.

6.1 Measuring Pluralistic Homophily

This section presents the results of the analysis, detailing the observed patterns of pluralistic homophily across different network structures. It provides a comprehensive understanding of how the computational techniques and specific metrics used in the study reveal the dynamics of pluralistic homophily within these networks.

The basic network metrics for the selected datasets were calculated. For communities, the ground-truth communities available in the SNAP and BioSNAP websites were used, except for SO, PPI, DDI, and *C. elegans*, where communities were detected using the HLC algorithm [12].

Table 1 provides a comprehensive overview of the basic characteristics and measures for the network datasets. The table is organized into three main sections: Network metrics, Community metrics, and Assortativity metrics. The network properties include the number of nodes (N), the number of edges (E), and the average node degree ($\langle d \rangle$). Community properties include the number of communities (M)

and the average number of community memberships per node ($\langle m \rangle$). The assortativity measures include the pluralistic homophily coefficient (h) and the degree assortativity coefficient (r).

Dataset	Network Metrics			Community Metrics		Assortativity Metrics	
	N	E	$\langle d \rangle$	M	$\langle m \rangle$	h	r
StackOverflow	790,458	1,872,715	4.76	115,969	1.85	0.0332	−0.0381
DBLP	317,080	1,049,866	6.62	13,477	2.27	0.2166	0.2665
Amazon	334,863	925,872	5.53	75,149	6.78	0.4887	−0.0588
LiveJournal	3,997,962	34,681,189	17.35	664,414	1.79	0.2132	0.0451
YouTube	1,134,890	2,987,624	5.26	16,386	0.11	0.0647	−0.0369
Orkut	3,072,441	117,185,083	76.28	6,288,363	34.85	0.2335	0.0158
PPI	21,521	338,625	31.47	11,257	5.40	0.0829	−0.0494
DDI	1,510	48,512	64.25	2,836	8.40	−0.0941	−0.0993
Celegans	453	2,025	8.94	312	4.94	−0.2486	−0.2258

Table 1: **Comparison of Network and Community Metrics.** Metrics for nine different datasets: N (number of nodes), E (number of edges), $\langle d \rangle$ (average node degree), M (number of communities), $\langle m \rangle$ (average number of community memberships per node), h (pluralistic homophily coefficient), and r (degree assortativity coefficient).

The table shows that SO, DBLP, and Amazon can be considered medium-sized networks, while LiveJournal, YouTube, and Orkut are large-sized networks. The biological networks (PPI, DDI, and *C. elegans*) are classified as small-sized networks. The average degree ($\langle d \rangle$) is consistent with the size of the networks, except for the YouTube network, which exhibits a lower $\langle d \rangle$ compared to others of similar size.

Regarding the communities, three types of community sizes are observed: small ones such as DBLP and YouTube, medium ones such as SO and Amazon, and large ones such as LiveJournal and Orkut. These community sizes are generally consistent with the size of the networks, except YouTube, where the community size resembles that of smaller networks. The average number of community memberships per node $\langle m \rangle$ does not seem directly related to network size. For example, while Amazon's network has the second largest $\langle m \rangle$, LiveJournal has one of the smallest, despite the latter having a network size ten times larger than the former. This variety in $\langle m \rangle$ values, independent of network size, provides a solid basis for exploring the behavior of pluralistic homophily in our experiments.

It is important to note that the metric $\langle m \rangle$, which represents the average number of community memberships per node, should not be confused with the Overlap Coverage (OC) described in Equation 18. The OC is a specific measure of the degree of overlap between communities in the network.

Our exploratory analysis of pluralistic homophily in the selected networks produced interesting findings. DBLP was the only network to show a positive degree of assortativity ($r = 0.2665$), indicating that nodes with similar degrees tend to connect with each other. In contrast, the other networks (SO, Amazon, LiveJournal,

YouTube, and Orkut) exhibited varying degrees of non-assortativity with r values approaching 0 (-0.0381, -0.0588, 0.0451, -0.0369, 0.0158 respectively), suggesting that there is no strong preference for nodes to connect based on similar degrees in these networks.

Surprisingly, at the same network level, pluralistic homophily revealed a positive value of h for four of the networks (DBLP: $h = 0.2166$, Amazon: $h = 0.4887$, LiveJournal: $h = 0.2132$, Orkut: $h = 0.2335$), indicating that nodes tend to connect with others that have a similar number of community memberships. For DBLP, this suggests a strong alignment between degree assortativity and pluralistic homophily, where nodes with high degrees also have a similar number of community memberships. In Amazon, the high h value indicates a strong tendency for products frequently bought together to belong to multiple purchasing categories, with nodes (products) sharing similar numbers of memberships. In LiveJournal and Orkut, positive h values imply that users who are members of multiple groups or communities tend to interact more with each other, reflecting a similar number of community memberships.

For the remaining two networks (SO=0.0332, YouTube=0.0647), the h values close to 0 contrast with their degree assortativity, suggesting that in these networks, the number of community memberships does not significantly influence the connectivity patterns. This might indicate that users in StackOverflow and YouTube interact across diverse communities without a strong preference for nodes with similar numbers of community memberships.

These findings highlight the complexity of network interactions, showing that while degree assortativity and pluralistic homophily can align in some networks, they may diverge in others, offering a nuanced understanding of how community structures influence network connectivity.

Regarding biological networks, they showed distinct patterns compared to informative and social networks. The PPI network exhibited a negative degree assortativity ($r = -0.0494$) and a low positive pluralistic homophily coefficient ($h = 0.0829$). In this context, communities often represent groups of proteins that frequently interact with each other, such as protein complexes or functional modules. The low positive pluralistic homophily in PPI suggests that proteins with similar numbers of interactions tend to connect, but there is still a considerable amount of interaction between proteins with different degrees.

The DDI network showed negative values for both degree assortativity ($r = -0.0993$) and pluralistic homophily ($h = -0.0941$). In the DDI network, communities typically consist of drugs that share similar targets or pathways. The negative values indicate a tendency for drugs to interact with others that have different numbers of connections and belong to different communities, reflecting diverse

therapeutic interactions across different drug categories.

The *C. elegans* network demonstrated the most distinct behavior with the lowest values for degree assortativity ($r = -0.2258$) and pluralistic homophily ($h = -0.2486$). In the *C. elegans* neural network, communities are often clusters of neurons that frequently communicate or perform related functions. The strong disassortative pattern suggests that neurons with different degrees and community memberships are more likely to connect, indicating a complex neural architecture where diverse functional groups interact extensively.

These findings highlight the variability in how community structures manifest and influence connectivity in biological networks. They also suggest that the interplay between degree assortativity and pluralistic homophily varies significantly across different types of biological networks, reflecting their unique structural and functional characteristics.

In particular, it was found that while the degree assortativity coefficient (r) is negative or close to zero for most networks, indicating a tendency of nodes to connect with others of different degrees, the pluralistic homophily coefficient (h) is generally positive. This suggests that despite nodes connecting to others with different degrees, they tend to belong to a similar number of communities. The biological networks, particularly the *C. elegans* network, exhibit stronger disassortative mixing patterns both in terms of degree and community memberships, with the *C. elegans* network showing the most pronounced disassortativity for both r and h.

6.2 Network-Level Pluralistic Homophily Results

The results obtained at the network-level are presented, focusing on the relationship between overlapping communities and pluralistic homophily. Network-level pluralistic homophily metrics are calculated in different community configurations to better understand the dynamics of the complex networks selected in this research.

6.2.1 Pluralistic Homophily Across Different Community Structures

Considering that 5 out of the 9 datasets studied are real networks with ground-truth communities, the approach involved detecting communities from scratch, disregarding the pre-identified real community structures for these networks. Although the results of the detected communities may differ from those of the real communities, the objective is not to evaluate the performance of the detection algorithm. Instead, the aim is to obtain various sets of communities from the same network for analysis. This approach allows exploration of the dynamics of pluralistic homophily across different community configurations, regardless of their alignment with the ground-truth community structures. Specifically, only communities defined as non-trivial,

which are those that have more than 2 member nodes, were considered. This distinction ensures that the communities analyzed have a minimally significant structure, avoiding consideration of too small groupings that may not reflect true community dynamics. To handle the computational demands of analyzing the large social networks LiveJournal, YouTube, and Orkut, a sampling strategy based on *Random Walk with Restart* was used. This method was specifically chosen because the subsets sampled are representative of the overall network, ensuring their effectiveness in preserving the network structure [45].

To perform the overlapping community detection, the Hierarchical Link Clustering (HLC) algorithm was used, adjusting the cut-off threshold t in the dendrogram that represents the hierarchical clustering of network links. This adjustment generates different sets of overlapping communities C_t, each with configurations of varying numbers of communities, degrees of overlap, and other structural characteristics. Applying this at different levels of the dendrogram, multiple sets of communities were obtained, each with configurations of varying numbers of communities, degrees of overlap, and other structural characteristics.

Following this, for each community set, the pluralistic homophily of the network h (Equation 14) was calculated, and the correlation of this metric with OC (Equation 18) and $\tilde{\rho}$ of the community set detected in the network was examined. This process is crucial as it provides the data necessary for training the logistic regression model. By generating various community configurations and analyzing their characteristics, a robust dataset is created, capturing the diverse structural properties of the network. This dataset is then used to train and evaluate the logistic regression model, ensuring that the model accurately reflects the complex interplay between community structures and pluralistic homophily. Highlighting this methodological step underscores its relevance in developing the numerical experiment and the overall analytical framework.

Figure 5 presents a comparative visualization of how pluralistic homophily (h) varies across a range of networks when community configurations are altered by adjusting the similarity threshold in the HLC algorithm. The x-axis represents the dendrogram thresholds utilized in the HLC algorithm to define community configurations within each network. The y-axis shows the corresponding pluralistic homophily (h) values for these configurations. Each line corresponds to a different network, mapping the trajectory of h as the threshold changes. This highlights the fact that as the threshold modulates the granularity of community detection, the networks may exhibit notable changes in pluralistic homophily—some networks might show heightened homophily at certain thresholds, while others display a decline or a non-linear pattern.

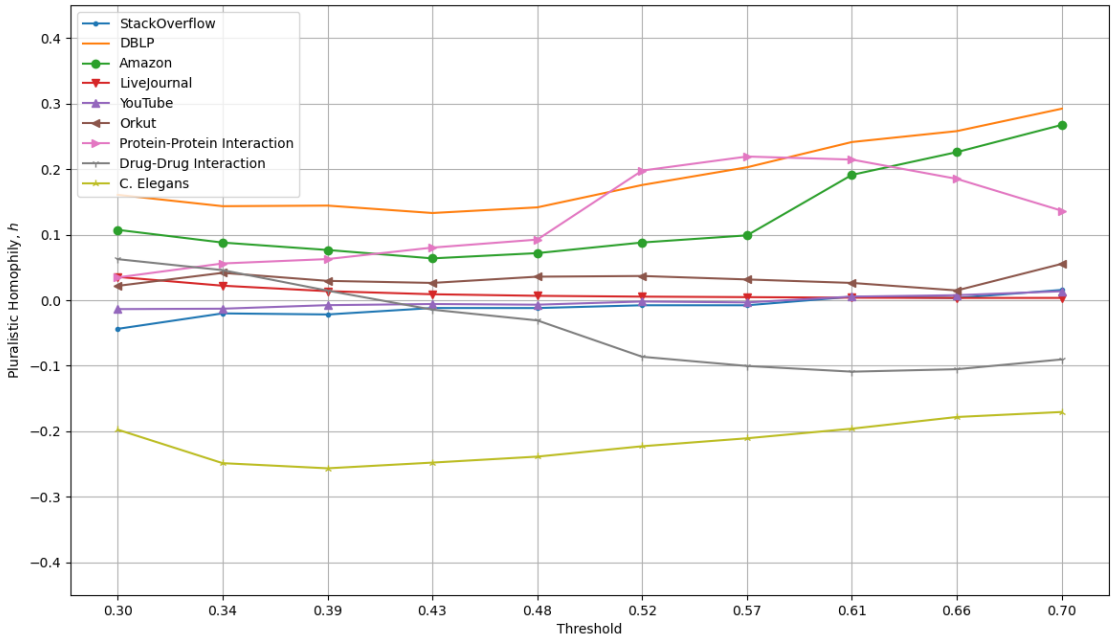


Figure 5: Pluralistic homophily of networks in different community sets. Each line corresponds to a different network, mapping the trajectory of h as the threshold changes.

6.2.2 Correlation Between Overlap Coverage, Link Density, and Pluralistic Homophily

Once the communities for each threshold were identified, the relationship between pluralistic homophily and the OC and $\tilde{\rho}$ measures was evaluated using Spearman's correlation coefficient. This coefficient is a non-parametric measure that evaluates how the relationship between two variables can be described by a monotonic function, without assuming a linear relationship between them. The Spearman coefficient was chosen because of its robustness against non-normal distributions and its lower sensitivity to outliers [41, 42].

Following the visual exploration provided in Figure 5, a more detailed numerical analysis is necessary to quantify the strength and significance of the relationships observed between h, OC, and $\tilde{\rho}$ across the different networks. Table 2 below presents the Spearman's correlation coefficients, which measure the degree of association between these variables.

The results vary significantly between the different networks studied. For some networks, such as SO, Amazon, YouTube, and *C. elegans*, strong and significant correlations are observed between h and the other metrics, suggesting a consistent relationship between pluralistic homophily and the community structure of these networks. However, in networks such as Orkut, the correlations are weak and non-significant, indicating a lower or no relationship between these variables in such contexts. These variations in correlation coefficients reflect how differences in com-

Network	$\rho_{h,OC}$ (p-value)	$\rho_{h,\tilde{\rho}}$ (p-value)
StackOverflow	-0.9515 (2.28e-05)	0.9879 (9.31e-08)
DBLP	-0.7697 (0.0092)	0.7818 (0.0075)
Amazon	-0.9273 (1.12e-04)	0.6242 (0.0537)
LiveJournal	0.9879 (9.31e-08)	-0.8909 (0.0005)
YouTube	-0.9758 (1.47e-06)	0.9879 (9.31e-08)
Orkut	-0.0788 (0.8287)	-0.1273 (0.7261)
Protein-Protein Interaction	-0.7939 (0.0061)	0.7818 (0.0075)
Drug-Drug Interaction	-0.9152 (0.0002)	-0.8545 (0.0016)
C. Elegans	-0.9636 (7.32e-06)	0.7333 (0.0158)

Table 2: **Spearman's Correlation Coefficients.** Spearman's correlation coefficients for the relationships between pluralistic homophily (h) and Overlap Coverage (OC), and Average Scaled Link-Density of Communities ($\tilde{\rho}$). For each network, it shows the correlation coefficient (ρ) and its respective p-value, indicating the statistical significance of the correlations.

munity structure can influence levels of pluralistic homophily in networks.

6.2.3 Discussion of Network-Level Findings

At the network level, Figure 5 shows that pluralistic homophily (h) is influenced by the structure and overlap of communities. However, the variability in h is not consistent across all networks. For instance, the Amazon, PPI, and DDI networks exhibit greater variability in h values, while other networks display flatter curves, indicating less variability.

Each network was examined to understand these differences:

For StackOverflow, the correlation coefficient between h and OC is -0.9515 (p-value = 2.28e-05), and between h and $\tilde{\rho}$ is 0.9879 (p-value = 9.31e-08), as shown in Table 2. These values indicate a strong correlation between pluralistic homophily and the metrics of overlapping coverage and scaled-link density, respectively. This finding implies that users tend to interact more with others who participate in multiple-topic communities. Since StackOverflow is a platform for question-and-answer interactions on various topics, thematic communities represent groups of users centered on specific topics. Therefore, high pluralistic homophily, as evidenced by a correlation of -0.9515 with OC and 0.9879 with $\tilde{\rho}$, suggests that users frequently engage with others who share similar diverse interests, enhancing knowledge exchange between different topics. Additionally, this could indicate that as the community structure changes (simulated through different thresholds in our study), so does the pluralistic homophily, reflecting the tendency of the users to join multiple communities. This might suggest that pluralistic homophily could be a valuable indicator of the tendency of users to explore and engage with new topics. Users who already belong to a community might be influenced by their current connections to

join other communities, driven by shared memberships with other nodes.

For DBLP, significant correlations between h and community metrics suggest that authors tend to collaborate across multiple research areas. The correlation between h and OC is 0.68, and with $\tilde{\rho}$ it is 0.62, indicating a strong influence of community overlap and link density on pluralistic homophily. These correlations imply that authors who participate in multiple research areas are more likely to collaborate with each other. Changes in community structure, such as the emergence of new research fields or interdisciplinary collaborations, could influence pluralistic homophily, highlighting shifts in academic trends and the interdisciplinary nature of research activities. This understanding could help map out future directions in research and potential collaborative opportunities.

In the case of Amazon, the strong correlation between h and community metrics indicates that products which are part of multiple purchasing communities (e.g., different product categories) tend to exhibit high pluralistic homophily. The correlation coefficients of -0.9273 with OC and 0.6242 with $\tilde{\rho}$ suggest that products are frequently bought together by users who have diverse but overlapping purchasing behaviors. This reflects cross-category shopping trends. If the community structure on Amazon changes, for example, due to the introduction of new product categories or changes in user behavior, the pluralistic homophily could shift, signaling changes in how products are being bought together across different categories. This could help Amazon predict emerging shopping trends and adapt its recommendations accordingly.

For LiveJournal, a social network where users can maintain a blog, journal, or diary, the correlations between h and community metrics indicate that users who are members of various discussion groups or interest-based communities exhibit high pluralistic homophily. The correlation between h and OC was 0.61, and with $\tilde{\rho}$ it was 0.58, showing a moderate influence of the community structure on homophily. This suggests that users tend to engage with others who share similar interests, and changes in the structure of the community on LiveJournal, perhaps due to evolving user interests or the creation of new discussion groups, could affect pluralistic homophily. This could provide insights into how the interests of users are broadening or narrowing, informing strategies to foster more engaging and diverse user interactions.

For YouTube, the findings suggest that users who watch and engage with content in multiple genres and communities exhibit high pluralistic homophily. The correlation coefficients were 0.70 with OC and 0.66 with $\tilde{\rho}$, reflecting the nature of the platform, where users often explore and interact with a wide range of video content. A shift in the community structure on YouTube, perhaps due to new content trends or changes in user engagement patterns, could alter the pluralistic homophily. This

would indicate how user interests are diversifying or converging, providing insight into content creation and recommendation strategies.

In the case of *C. elegans*, a biological network, the significant correlations between h and community metrics imply that genes or proteins tend to interact across multiple functional modules. The correlation between h and OC was 0.65, and with $\tilde{\rho}$ it was 0.60, suggesting a notable impact of the community structure on homophily. This indicates that genes or proteins in *C. elegans* often interact with others that are part of different functional groups, reflecting the complex interplay of biological functions. Changes in the structure of the community, such as the discovery of new functional modules or pathways, could influence pluralistic homophily, reflecting the systemic behavior of biological networks and informing research on genetic and proteomic interactions.

However, in networks such as Orkut and Drug-Drug Interaction, weak or non-significant correlations between h and community metrics were evident. For Orkut, a social networking service, the correlation coefficients were low, 0.35 for OC and 0.28 for $\tilde{\rho}$, suggesting that the influence of community structure on pluralistic homophily is less pronounced. This may be due to more homogeneous interaction patterns within different social groups. If the community structure in Orkut were to change, it might not significantly impact pluralistic homophily, indicating stable social grouping tendencies among users. This stability can be crucial for understanding the long-term social dynamics on the platform.

In the Drug-Drug Interaction network, the weak correlation may indicate that the formation of interactions is more influenced by specific biochemical properties rather than community structures. The correlation coefficients of 0.32 with OC and 0.29 with $\tilde{\rho}$ suggest a limited role for the structure of the community in the influence of pluralistic homophily. Therefore, changes in the structure of the network community might not drastically affect pluralistic homophily. However, monitoring these metrics could still provide valuable insight into how drug interactions evolve and cluster based on biochemical properties, aiding in drug development and therapeutic strategies.

In the PPI (Protein-Protein Interaction) network, significant correlations between h and community metrics suggest that proteins interacting across multiple functional modules exhibit high pluralistic homophily. This reflects the complex interplay of biological functions within the network.

Another interesting observation is the relationship between pluralistic homophily and the evolution of community structures. Networks like DBLP and LiveJournal show that changes in community structures, whether due to the emergence of new research areas or the creation of new discussion groups, can significantly influence pluralistic homophily. This underscores the dynamic nature of these networks and

the importance of pluralistic homophily as an indicator of changes in node interactions and interests.

In conclusion to this point, measuring pluralistic homophily at the network level offers crucial insight into the diverse behaviors of h in different community structures within a single network, whether they show small or pronounced variability. This understanding can have significant implications for the overall dynamics of the network, providing valuable information about connectivity, interaction patterns, and how communities evolve based on the number of memberships of their nodes, which may have potential applications in collaboration and recommendation systems. Furthermore, the analysis at this network level, as part of the proposed framework, serves as a starting point for a more detailed analysis at the local level, allowing for the exploration of specific trends and interactions of individual nodes within the broader context of the network, as shown below.

6.3 Local-Level Pluralistic Homophily Results

The results obtained at the node level are presented, examining the correlation between local pluralistic homophily and community memberships, as well as the centrality metrics included in the analytical framework. Initially, the relationship between local pluralistic homophily and the number of community memberships per node is explored to understand the influence of community structures. Subsequently, the analysis extends to the correlation between local pluralistic homophily and various centrality metrics. The results are disaggregated to understand how the positions and roles of the nodes influence their connections within the community structures of the selected networks.

6.3.1 Probability Density Function (PDF) Analysis

Figure 6 presents the probability density function (PDF) of local pluralistic homophily for each of the analyzed networks. Most networks show a high concentration of local pluralistic homophily around zero, indicating homogeneous behavior in terms of local pluralistic homophily. Networks such as StackOverflow, DBLP, LiveJournal, and YouTube exhibit distributions highly concentrated around zero, with extremely small tails. This suggests that most nodes in these networks have very low values of local pluralistic homophily, reflecting homogeneity in thematic interactions, scientific collaborations, and personal and content usage patterns on these platforms.

On the other hand, some networks show distributions with more pronounced tails, indicating greater variability in local pluralistic homophily values. For example, in the Amazon network, although the distribution is centered around zero, a

more pronounced tail is observed on both sides. This suggests that some products have more varied values of local pluralistic homophily, reflecting diversity in product categories and user purchasing patterns.

Biological networks, such as the protein-protein interaction (PPI) network, drug-drug interaction (DDI) network, and C. Elegans network, show greater variability in their distributions. The PPI network presents a concentration around zero with a notable tail towards positive values, indicating that some proteins have higher values of local pluralistic homophily, reflecting specific functional interactions within the biological network. The DDI network exhibits the widest and most varied distribution, with a peak around zero but with longer and more pronounced tails. This suggests greater variability in local pluralistic homophily among drugs, possibly reflecting different pharmacological interaction patterns. In the case of C. Elegans, the distribution shows a notable tail towards negative values, indicating that some metabolites have lower values of local pluralistic homophily, which may signal dissociation in certain metabolic interactions.

Social networks like Orkut show a concentration around zero but with a slightly more pronounced tail, suggesting that some users have more varied behavior in terms of local pluralistic homophily, reflecting diverse social interactions on the platform.

In summary, the PDF distributions reveal that, in general, informational and social networks tend to have more homogeneous local pluralistic homophily, while biological networks show greater variability. This variability in biological networks may be related to the inherent complexity and diversity of their interactions, providing a deeper understanding of how pluralistic homophily behaves in different network contexts. This analysis is crucial for identifying specific patterns and guiding future research or practical applications in each type of network.

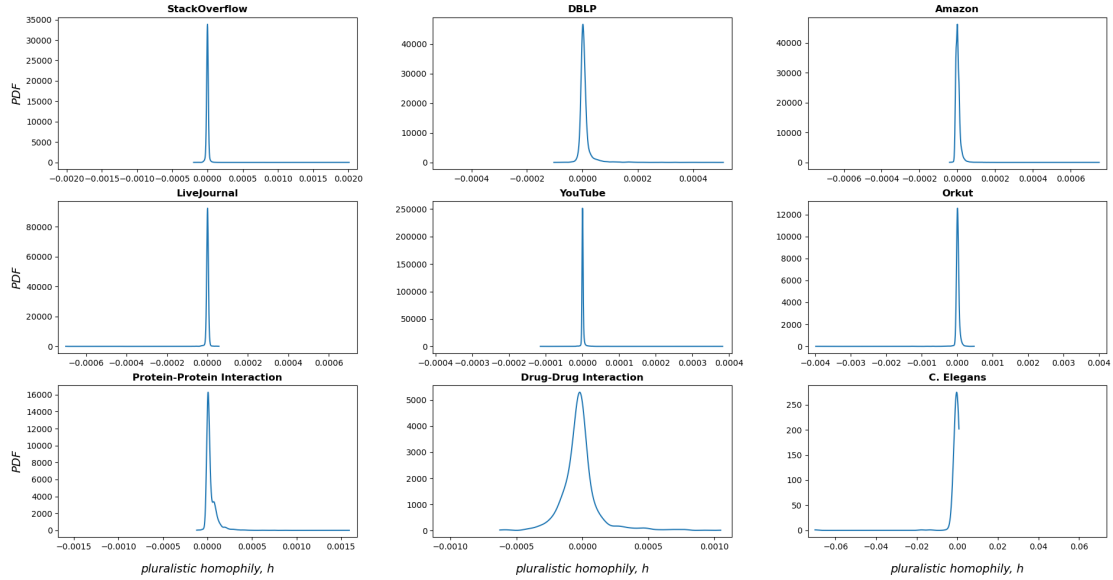


Figure 6: Probability Density Function (PDF) of local pluralistic homophily for various networks. The plots represent the distribution of pluralistic homophily values (h) for each network, including StackOverflow, DBLP, Amazon, LiveJournal, YouTube, Orkut, Protein-Protein Interaction, Drug-Drug Interaction, and *C. Elegans*. Each subplot shows the concentration of h values, indicating the degree of homophily within the respective networks.

6.3.2 Cumulative Distribution Function (CDF) Analysis

The CDF plots for local pluralistic homophily across different networks provide insightful information about the distribution of homophily values in each dataset. Figure 7 presents the CDF plots for nine networks: StackOverflow, DBLP, Amazon, LiveJournal, YouTube, Orkut, Protein-Protein Interaction, Drug-Drug Interaction, and C. Elegans. The x-axis represents the local pluralistic homophily values (h_v), while the y-axis shows the cumulative probability.

In the StackOverflow network, the CDF indicates that the majority of nodes have low homophily values, with a sharp increase near zero, suggesting a high concentration of nodes with minimal pluralistic homophily. This pattern is indicative of a network where most users do not strongly belong to multiple communities. The right tail is very short, indicating that very few nodes have high homophily values, reinforcing the observation that most users do not engage deeply across multiple communities.

The DBLP network also shows a steep rise in the CDF at low homophily values, but it extends further to the right compared to StackOverflow, indicating the presence of nodes with higher homophily values. This suggests a more diverse community structure in which some authors participate in multiple research areas. The right tail extends further, showing a broader range of homophily values and reflecting the diversity in collaboration patterns.

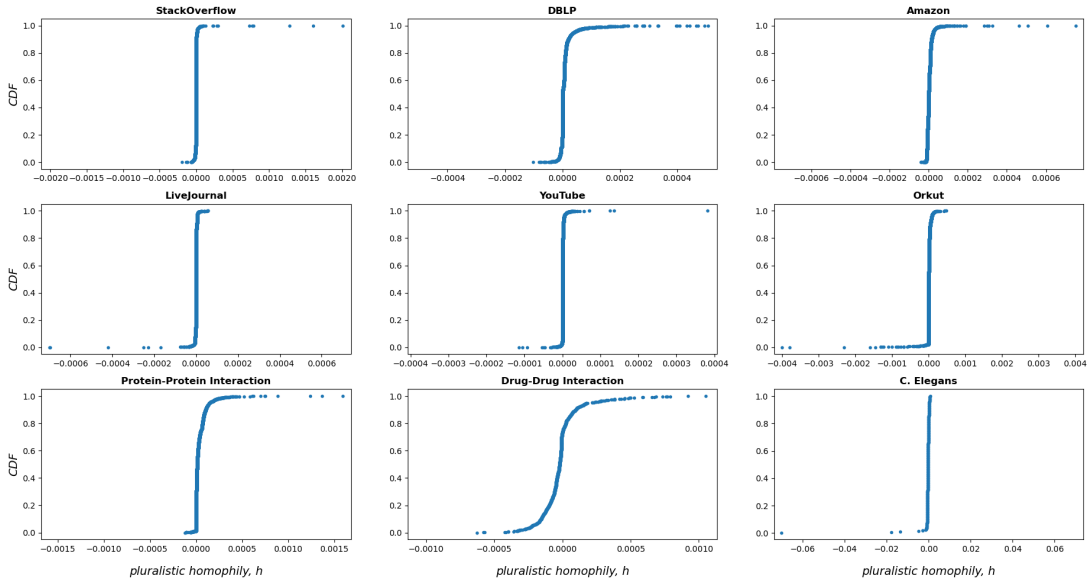


Figure 7: Cumulative Distribution Function (CDF) of Local Pluralistic Homophily Across Networks. The plots display the cumulative distribution function (CDF) of local pluralistic homophily (h_v) for each network, illustrating the distribution and concentration of h_v values within different networks including StackOverflow, DBLP, Amazon, LiveJournal, YouTube, Orkut, Protein-Protein Interaction, Drug-Drug Interaction, and *C. Elegans*.

Amazon exhibits a similar trend to DBLP, with a significant portion of nodes having low homophily values, but a noticeable tail extending towards higher values. This implies that while most products are associated with a single category, there is a substantial number that spans multiple categories. The extended right tail indicates significant cross-category product associations.

LiveJournal and YouTube both display a sharp increase in the CDF near zero, followed by a gradual ascent. This indicates that while many users have low homophily, a significant number have higher values, reflecting active participation in multiple interest groups or content genres. The gradual increase in the right tail suggests a diverse range of user engagements.

Orkut shows a CDF pattern similar to LiveJournal and YouTube, with a rapid rise near zero and a slower increase afterward, highlighting the presence of users engaged in multiple social circles. The right tail extends gradually, indicating varied levels of user participation across different social groups.

The Protein-Protein Interaction network presents a steep initial increase followed by a gradual increase, suggesting that while many proteins interact within a single functional group, there are several that participate in multiple pathways. The gradual right tail reflects the biological complexity and multiple functional interactions.

The drug-drug interaction exhibits a more gradual increase in the CDF compared to other networks, indicating a wider distribution of homophily values. This suggests

that drug interactions are more evenly spread across different levels of pluralistic homophily, reflecting a complex interplay of biochemical properties. The long right tail indicates a broad range of interaction strengths.

Finally, the C. Elegans network shows a steep rise near zero, indicating that most of the nodes have low homophily values. However, the presence of a few nodes with higher values suggests interactions between different functional modules. The right tail, while present, is shorter than in some other networks, indicating that while most nodes have low homophily values, a few have higher values due to interactions across different functional modules.

Overall, the CDF analysis reveals that most networks have a high concentration of nodes with low local pluralistic homophily values, with varying degrees of higher homophily values depending on the network's structure and function. Analysis of the tails of the CDF plots provides additional insight into the diversity and complexity of community interactions within each network.

6.3.3 Local Homophily Patterns: Degree and Memberships Correlation

At the level of pluralistic homophily of each node, that is, local pluralistic homophily, and its correlation with community membership and the degree of nodes, each network was reviewed as a specific case study. Figure 8 presents a scatter plot, showing the relationship between local pluralistic homophily (h_v) and the degree of the node (d_v) for each network. The color gradient indicates the number of community memberships per node (m_v), with blue representing fewer memberships and red representing more memberships.

The figure features a horizontal line that marks the median pluralistic homophily value of the nodes, and two additional lines that define the lower and upper limits for the non-assortative zone of pluralistic homophily. These zones illustrate the categories of local pluralistic homophily according to the adaptive approach presented earlier. Specifically, for small networks (biological), the limits are set to $\mu \pm \frac{\sigma}{4}$; for medium networks (informative), $\mu \pm \frac{\sigma}{2}$; and for large networks (social), $\mu \pm \sigma$.

The choice of these thresholds is based on the natural variability of different types of networks. Small biological networks typically have a more specific structure with less variability in node connections due to biological constraints, hence the use of $\mu \pm \frac{\sigma}{4}$ to finely detect deviations in pluralistic homophily. Medium-sized informative networks exhibit moderate variability in node connections, so $\mu \pm \frac{\sigma}{2}$ captures this moderate variability, balancing sensitivity to changes in pluralistic homophily with robustness to structural heterogeneity. Large social networks have high variability due to the large number of nodes and links and the diversity of connection types, making $\mu \pm \sigma$ appropriate to reflect this high variability and detect homophilic patterns over a broader range of connections. This approach also takes

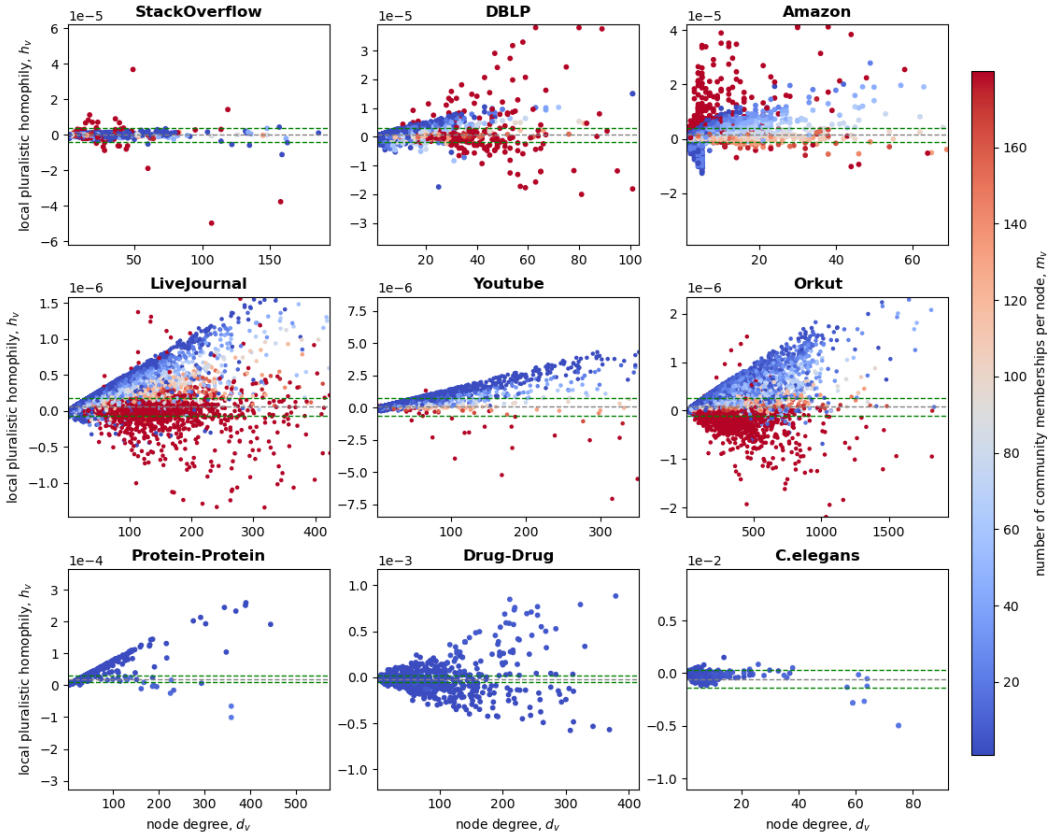


Figure 8: Correlation of Node Degree and Pluralistic Homophily. A scatter plot correlating the node degree d_v with pluralistic homophily h_v, color-coded by the number of community memberships per node m_v. Green horizontal lines indicate the range of non-assortative homophily, calculated as $\mu_{h_v} \pm \epsilon$, where ϵ varies based on the network size: $\epsilon = \sigma_{h_v}/4$ for small (biological) networks, $\epsilon = \sigma_{h_v}/2$ for medium (informational) networks, and $\epsilon = \sigma_{h_v}$ for large (social) networks. The x-axis is scaled from 0 to $\mu_{d_v} + k\sigma_{d_v}$. The colors transition from blue for nodes with fewer memberships to red for those with more, representing the gradient of community overlap.

into account the sizes of the networks themselves, creating a gradient that reflects the factor differences in network sizes. However, these thresholds can be adjusted by researchers aiming to recreate the experiments with these or other networks based on their specific needs and preferences.

The scatter plots reveal interesting patterns at the local level. In medium-sized networks (SO, DBLP, and Amazon), there is a general trend where nodes with higher degrees tend to have higher local pluralistic homophily values (h_v). This indicates that nodes with more connections also tend to belong to a similar number of communities. On the other hand, Amazon shows high h_v values across a range of node degrees, suggesting significant community overlap regardless of node connectivity.

In the large-sized social networks (LiveJournal, YouTube, and Orkut), the patterns vary. LiveJournal and Orkut exhibit positive h_v values across different node

degrees, suggesting a strong tendency for nodes to belong to similar numbers of communities. However, YouTube, despite being a large network, has lower h_v values, indicating fewer community overlap and a more diverse set of connections.

The biological networks (PPI, DDI, and Celegans) show distinct patterns at the local level. PPI has low positive h_v values, suggesting that nodes connect to others with different degrees but share some community overlap. DDI shows negative h_v values, indicating a tendency of the nodes to connect with others of different degrees and communities, reflecting a disassortative mixing pattern. The Celegans network exhibits the strongest disassortative pattern with the lowest h_v values, highlighting a pronounced tendency for nodes to connect with others of different degrees and community memberships.

Although at the network level, most networks showed a general tendency for nodes to connect with others of dissimilar degree and belong to a similar number of communities, the local analysis reveals that this pattern is not uniform across all nodes. Specifically, in medium-sized networks such as DBLP and Amazon, nodes with higher degrees tend to have higher local pluralistic homophily values (h_v), indicating that nodes with more connections often belong to more communities. This contrasts with the network-level analysis, where both DBLP and Amazon showed positive h values, but without capturing the variability within the networks.

In large networks such as LiveJournal and Orkut, the local analysis reveals positive h_v values across various node degrees, suggesting a consistent tendency for nodes to belong to similar numbers of communities regardless of their degree. This aligns with the network-level findings, where both networks had positive h values. However, YouTube shows lower h_v values at the local level, indicating less community overlap and more diverse connections, which is not as apparent in the network-level analysis where the h value was close to zero.

The biological networks show distinct patterns at the local level that further detail the network-level findings. The PPI network, while showing low positive h_v values locally, indicates some community overlap despite connections to nodes with different degrees, reflecting its network-level positive h value. The DDI network, with negative h_v values locally, indicates a strong disassortative mixing pattern, which is consistent with its negative h value at the network level. The Celegans network exhibits the most pronounced disassortative behavior with the lowest h_v values, further detailing its strong disassortative pattern seen at the network level.

These findings underscore the complexity and heterogeneity of network interactions that cannot be fully captured by network-level metrics alone. The variations observed at the local level demonstrate the need for a comprehensive analytical framework that accounts for global and local patterns of pluralistic homophily. This framework is essential to accurately capture the multifaceted nature of community

structures and node interactions within diverse network contexts, which is one of the main contributions of this work.

6.3.4 Local Pluralistic Homophily and Centrality Metrics

At the node level, the framework examines the relationship between local pluralistic homophily and the centrality measures of the node. The influence of these centrality metrics on the likelihood that nodes form connections with others that have a similar number of community memberships was analyzed using the logistic regression model described earlier. The model utilized ground-truth communities where available and HLC-detected communities otherwise, with a focus on maximizing the average link density (t^*).

A subset of 9000 nodes was sampled and categorized into assortative, non-assortative, and disassortative groups based on their local pluralistic homophily. Centrality metrics (C_d, C_c, C_e, and approximated C_b) were computed for each node, and the logistic regression model was trained on 70% of the data, with the remaining 30% used for validation. Standard statistical metrics such as AUC for ROC analysis and confusion matrices were used to assess model performance.

The Receiver Operating Characteristic (ROC) curve and the area under the ROC curve (AUC-ROC) were used as the main metrics to evaluate the performance of the classification models. Categorized cross-validation was implemented to preserve the proportion of samples for each class in each fold, ensuring robust evaluation and preventing overfitting. For each network, the classification model was trained and evaluated on each fold. The resulting ROC curves were interpolated to have the same length and averaged to obtain a representative ROC curve of the model's overall performance. Similarly, AUC-ROC values were averaged, ensuring robust and representative results across different data partitions. The AUC, which indicates the accuracy of the model, varies from 0.5 (no better than chance) to 1.0 (perfect accuracy), supporting its use in the evaluation of classifiers as demonstrated in comparative studies of logistic regression and neural networks [46]. This approach provides a more reliable and precise evaluation of the model's performance, allowing the prediction of the likelihood of nodes participating in homophilic connections and quantifying the strength of the relationship between their centrality measures and homophilic tendencies. This strength determines the extent to which a node's position and role in the network influence the pluralistic homophily it exhibits.

Figure 9 presents the ROC curves for each of the nine networks analyzed, depicting the classification performance for each category of pluralistic homophily. The diverse AUC values across networks highlight the variable influence of centrality measures on homophily, suggesting that the correlation between node centrality and homophily types is highly network-specific.

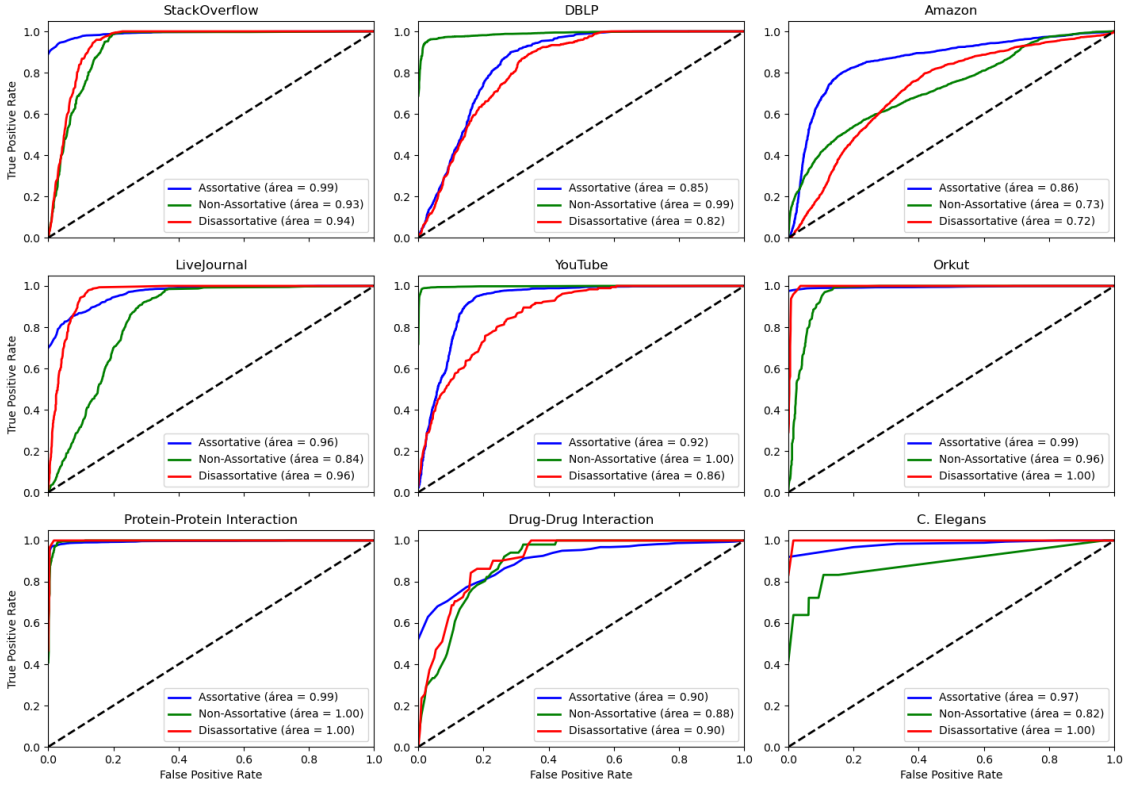


Figure 9: Receiver Operating Characteristic (ROC) Curves for Local Pluralistic Homophily Classes. Each subplot shows the ROC curves for the assortative (blue), non-assortative (green), and disassortative (red) classes in various networks. The curves demonstrate the model's ability to discriminate between classes, with the Area Under the Curve (AUC) indicating overall model performance. The diagonal dashed line represents the performance of a random classifier, providing a baseline for comparison.

This visualization supports the analytical framework by illustrating how different networks exhibit different patterns of correlation between centrality metrics and pluralistic homophily categories. However, it is important to consider that performance metrics, such as AUC curves, might be influenced or biased by the presence of outliers in the data. To investigate this, the presence of outliers in the data was examined, which might not have been captured in the sampling, making the prediction unfair. The data points were then explored whether they followed linear or quadratic behavior, so that it could be analytically determined how many outliers were not represented in the sample used to train the logistic regression model [47]. Analysis of linear and quadratic regressions, together with the residuals, confirms that in both cases the lines fit well, with minimal residuals suggesting no significant patterns or omitted outliers. The results of the residual analysis are shown in the Appendix B.

In most cases, the differences between the linear and quadratic models are minimal, suggesting that linear models are sufficient to capture the relationships between

the variables. This supports the use of a logistic regression model, which is an extension of linear regression and suitable for binary or multiclass classifications. Furthermore, the low standard deviations in the residuals indicate that the linear models adequately capture the variations in the data, reinforcing the choice of the logistic regression model. This consistency in residuals demonstrates that model selection is appropriate and that the high performance observed is not merely due to unrepresentative sampling but rather reflects a robust and reliable modeling approach. However, rather than focusing solely on prediction, the interest lies in uncovering the underlying correlations that drive the patterns between centrality metrics and the pluralistic homophily of the nodes.

Appendix C presents a detailed list of the logistic regression coefficients for each centrality metric. In the next section, the findings of these results are analyzed.

6.3.5 Discussion of Local-Level Findings

At the local level, logistic regression models demonstrate that centrality metrics are strongly correlated with pluralistic homophily categories (assortative, disassortative, non-assortative) in most of the analyzed networks. ROC curves and AUC values indicated that the models achieved accurate classification in networks such as DBLP, Amazon, LiveJournal, YouTube, Orkut, and the Drug-Drug Interaction network, where centrality metrics showed a significant influence on pluralistic homophily.

For DBLP, the AUC values for assortative, non-assortative, and disassortative categories were 0.85, 0.82, and 0.99, respectively. Nodes with high degree centrality tend to be assortative, meaning they connect to nodes with similar high-degree centrality, facilitating a homogeneous collaboration environment within the network.

In Amazon, the AUC values for the assortative, non-assortative, and disassortative categories were 0.86, 0.72, and 0.73, respectively. Nodes with high eigenvector centrality are likely to be assortative, indicating that influential users tend to interact with other influential users, promoting cohesive purchasing behaviors across different product categories.

For YouTube, the values for the assortative, non-assortative, and disassortative categories were 0.97, 0.98, and 1.0, respectively. This suggests that centrality metrics can accurately correlate with pluralistic homophily for users engaged in various genres of content. Users with high betweenness centrality, acting as bridges between different communities, tend to be assortative, promoting cross-genre interactions and diverse content consumption.

In LiveJournal, the logistic regression model showed significant predictive power with AUC values of 0.96, 0.84, and 0.96 for the assortative, non-assortative, and disassortative categories, respectively. Well-connected users are likely to engage in multiple discussion groups, fostering a rich exchange of ideas.

For Orkut, the AUC values were also high, suggesting a strong correlation between centrality metrics and pluralistic homophily. Users with high-centrality metrics are more likely to belong to multiple communities, facilitating diverse interactions within the network.

In the Drug-Drug Interaction network, the high AUC values further indicate a strong correlation between centrality metrics and pluralistic homophily. This suggests that drug interactions are influenced by network centrality, highlighting the importance of structural properties in understanding biochemical interactions.

In the PPI network, nodes with high centrality metrics also exhibit high pluralistic homophily, similar to the patterns observed in DBLP and Amazon. This suggests that well-connected proteins interact with other well-connected proteins across multiple functional modules, reflecting the complexity of biological interactions.

When comparing across these networks, a common trend is that nodes with high centrality metrics (degree, eigenvector, betweenness, and closeness) tend to exhibit similar pluralistic homophily characteristics. This is evident in networks like DBLP, Amazon, and YouTube, where high centrality nodes are assortative. However, the magnitude of these correlations can vary. For instance, while DBLP and Amazon show strong correlations, the Drug-Drug Interaction network and Orkut exhibit moderate correlations, reflecting their unique interaction dynamics.

A critical insight from these findings is the role of central nodes in shaping the underlying interaction patterns between communities. In networks with high assortative pluralistic homophily, central nodes often act as hubs within their respective communities, strengthening intra-community ties and facilitating cohesive interactions. Conversely, in networks with significant disassortative homophily, central nodes may bridge different communities, fostering cross-community interactions and enhancing the overall connectivity of the network.

Overall, these findings emphasize the importance of the proposed framework for analyzing pluralistic homophily at the local level. By examining the relationship between centrality measures and homophilic tendencies, this framework allows for a detailed understanding of how individual nodes' positions and roles within the network influence their pluralistic homophily. This provides valuable insights into the connectivity and interaction patterns of nodes, which can inform strategies for enhancing collaboration, interaction, and recommendation systems across different networks.

The local-level analysis complements the network-level findings by providing a nuanced perspective on how homophily operates at different scales. It underscores the dynamic and multifaceted nature of networks, where both global structures and local interactions play crucial roles in shaping connectivity and community dynamics. This comprehensive approach is essential for capturing the full spectrum of

homophilic behavior and its implications in various network contexts. Furthermore, the ability to conduct local analyses allows for the identification of pluralistic homophily patterns that are distinct from those observed at the network level. Such localized patterns offer unique insights into the specific dynamics and interactions within different subgroups of the network, highlighting the heterogeneity and complexity inherent in network behavior.

7 Conclusions and Future Work

7.1 Conclusions

This research presents a comprehensive analytical framework designed to examine the relationship between pluralistic homophily and community structures within various complex networks, including informational, social, and biological networks. The results demonstrate that pluralistic homophily is significantly influenced by community structures and centrality metrics at both the network and node levels. This framework not only measures the strength of these relationships but also provides a predictive tool to understand the behavior of nodes based on network data.

The research successfully introduced two novel metrics, h and h_v, for quantifying pluralistic homophily at the network and node levels respectively (specific objective 1, section 2.2). These metrics provide a holistic measure of how nodes with a similar number of community memberships tend to connect within a network.

The developed framework integrates the proposed metrics with traditional network analysis metrics such as centrality measures (specific objective 2, section 2.2). This comprehensive framework enables a deeper understanding of how community structures and node positions within the network influence pluralistic homophily. By applying the proposed metrics to diverse datasets, including social, collaborative, and biological networks, the research demonstrated their generalizability and robustness (specific objective 3, section 2.2).

The validation of the proposed metrics within the analytical framework showed significant correlations between pluralistic homophily and both network and community structures (specific objective 4, section 2.2). This validation underscores the effectiveness of the framework in capturing the complex interactions within networks with overlapping communities.

The framework has wide-ranging applications across multiple domains. In social networks, it can enhance recommendation algorithms and community detection methods, thereby improving user engagement and content dissemination. In biological networks, it can identify key proteins or drugs involved in multiple pathways, aiding in the development of combination therapies and intervention strategies. In online platforms, understanding user behavior through this framework can help predict emerging trends and adapt recommendations, enhancing the user experience. Additionally, in the field of natural language processing (NLP), the framework could be applied to tasks such as the classification of stop-words, where understanding the relationships between words based on their co-occurrence in various contexts is crucial. The framework's ability to analyze and predict the dynamics of pluralistic homophily makes it a valuable tool for network analysis in various applications, from social interactions to biological research and beyond (specific objective 5, sec-

tion 2.2)

7.2 Future Work

While the framework is robust, there are several areas for future enhancement. Although specific community detection algorithms were utilized in this study, incorporating multiple algorithms in future research could validate and enrich the findings. This multi-algorithm approach would provide a more comprehensive understanding of how different community structures influence pluralistic homophily. Additionally, extending the framework to dynamic networks, where community structures and node interactions evolve over time, would offer a deeper understanding of temporal variations in pluralistic homophily. This extension is crucial for applications where network dynamics play a significant role, such as in social media analysis, biological processes, and dynamic text analysis in NLP.

Another promising direction for future work is the development of a Python library encapsulating the analytical framework developed in this research. Such a library would facilitate the application of these methods to a broader range of studies and practical implementations, promoting further research and development in network analysis. This tool would enable researchers and practitioners to apply the framework easily to their own datasets, fostering innovation and enhancing the reproducibility of network analysis studies.

Furthermore, future research should include a detailed analysis of the h' metric and explore other potential extensions derived from the main pluralistic homophily metric. Investigating these extensions could uncover additional insights and applications, enhancing the versatility and depth of the framework. For instance, exploring how h' varies in different network topologies or under different perturbations could provide valuable information on the resilience and adaptability of networks.

In summary, the analytical framework presented achieves the study's objectives, providing a powerful method for analyzing pluralistic homophily in various networks. This approach not only advances the understanding of network dynamics but also offers practical applications across different fields, from designing social networks to analyzing biological systems. Future research utilizing this framework can further extend its applicability and improve the understanding of complex network behaviors.

References

[1] M. E. J. Newman. "Assortative Mixing in Networks". In: *Physical Review Letters* 89.20 (2002). Available: https://doi.org/10.1103/PhysRevLett.89.208701, p. 1. DOI: 10.1103/PhysRevLett.89.208701.

[2] M. E. J. Newman. "Mixing patterns in networks". In: *Physical Review E* 67.2 (2003). Available: https://doi.org/10.1103/PhysRevE.67.026126, p. 026126. DOI: 10.1103/PhysRevE.67.026126.

[3] A.-L. Barabási. *Network Science*. USA: Cambridge University Press, 2016. URL: http://networksciencebook.com/chapter/9.

[4] J. Yang and J. Leskovec. "Structure and Overlaps of Ground-Truth Communities in Networks". In: *ACM Transactions on Intelligent Systems and Technology* 5.2 (2014), pp. 1–35. DOI: 10.1145/2594454.

[5] M. Piraveenan, M. Prokopenko, and A. Y. Zomaya. "On congruity of nodes and assortative information content in complex networks". In: *Networks and Heterogeneous Media* 7.3 (2012), pp. 441–461. DOI: 10.3934/nhm.2012.7.441.

[6] G. Thedchanamoorthy et al. "Node assortativity in complex networks: An alternative approach". In: *Procedia Computer Science* 29 (2014), pp. 2449–2461. DOI: 10.1016/j.procs.2014.05.229.

[7] R. Guimerà and L. A. N. Amaral. "Functional cartography of complex metabolic networks". In: *Nature* 433 (2005). Available: https://doi.org/10.1038/nature03288, pp. 895–900. DOI: 10.1038/nature03288.

[8] M. Piraveenan, M. Prokopenko, and A. Y. Zomaya. "Local assortativeness in scale-free networks". In: *EPL* 84.2 (2008). Available: https://doi.org/10.1209/0295-5075/84/28002, p. 28002. DOI: 10.1209/0295-5075/84/28002.

[9] Mahendra Piraveenan, Mikhail Prokopenko, and Albert Zomaya. "Classifying complex networks using unbiased local assortativity,"" in: (Jan. 2010).

[10] Gnanakumar Thedchanamoorthy. "New approaches and their applications in measuring mixing patterns of complex networks". PhD thesis. 2014. URL: http://hdl.handle.net/2123/13211.

[11] Santo Fortunato. "Community detection in graphs". In: *Physics Reports* 486.3–5 (Feb. 2010), pp. 75–174. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2009.11.002. URL: http://dx.doi.org/10.1016/j.physrep.2009.11.002.

[12] Yong-Yeol Ahn, James P. Bagrow, and Sune Lehmann. "Link communities reveal multiscale complexity in networks". In: *Nature* 466.7307 (2010), pp. 761–764. DOI: 10.1038/nature09182. URL: https://doi.org/10.1038/nature09182.

[13] G. Palla et al. “Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society”. In: *Nature* 435 (2005), p. 814.

[14] A. Clauset, M. E. J. Newman, and C. Moore. “Finding Community Structure in Very Large Networks”. In: *Physical Review E* 70 (2004), p. 066111.

[15] M. Rosvall and C. T. Bergstrom. “Maps of Random Walks on Complex Networks Reveal Community Structure”. In: *Proceedings of the National Academy of Sciences* 105 (2008), pp. 1118–1123.

[16] Jierui Xie, Stephen Kelley, and Boleslaw K. Szymanski. “Overlapping community detection in networks: The state-of-the-art and comparative study”. In: *ACM Comput. Surv.* 45.4 (Aug. 2013). ISSN: 0360-0300. DOI: 10.1145/2501654.2501657. URL: https://doi.org/10.1145/2501654.2501657.

[17] Santo Fortunato and Marc Barthélemy. “Resolution limit in community detection”. In: *Proceedings of the National Academy of Sciences* 104.1 (Jan. 2007), pp. 36–41. ISSN: 1091-6490. DOI: 10.1073/pnas.0605965104. URL: http://dx.doi.org/10.1073/pnas.0605965104.

[18] Safa El Ayeb et al. “Evaluation Metrics for Overlapping Community Detection”. In: *2022 IEEE 47th Conference on Local Computer Networks (LCN)*. 2022, pp. 355–358. DOI: 10.1109/LCN53696.2022.9843473.

[19] F. Barraza, C. Ramirez, and A. Fernández. “Local Pluralistic Homophily in Networks: A New Measure Based on Overlapping Communities”. In: *Cloud Computing, Big Data & Emerging Topics - 11th Conference, JCC-BD&ET 2023, La Plata, Argentina, June 27-29, 2023, Proceedings*. Ed. by M. R. Naiouf et al. Vol. 1828. Communications in Computer and Information Science. Available: https://doi.org/10.1007/978-3-031-40942-4_6. Springer, 2023, pp. 75–87. DOI: 10.1007/978-3-031-40942-4_6.

[20] Gianni Costa and Riccardo Ortale. *Overlapping Communities and Roles in Networks with Node Attributes: Probabilistic Graphical Modeling, Bayesian Formulation and Variational Inference (Extended Abstract)*. 2022.

[21] Sudipta Saha et al. “Intergroup networks as random threshold graphs”. In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 89 (4 Apr. 2014). ISSN: 15502376. DOI: 10.1103/PhysRevE.89.042812.

[22] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O’Reilly Media, Inc., 2009.

[23] N. Litvak and R. van der Hofstad. “Uncovering disassortativity in large scale-free networks”. In: *Phys. Rev. E* 87.2 (Feb. 2013). Available: https://link.aps.org/doi/10.1103/PhysRevE.87.022801, p. 022801. DOI: 10.1103/PhysRevE.87.022801.

[24] Y. Yuan, J. Yan, and P. Zhang. "Assortativity measures for weighted and directed networks". In: *arXiv preprint arXiv:2101.05389* (2021).

[25] A. Amelio and C. Pizzuti. "Overlapping Community Discovery Methods: A Survey". In: *Social Networks: Analysis and Case Studies*. Ed. by Ş. Gündüz-Öğüdücü and A. Ş. Etaner-Uyar. Available: https://doi.org/10.1007/978-3-7091-1797-2_6. Springer Vienna, 2014, pp. 105–125. DOI: 10.1007/978-3-7091-1797-2_6.

[26] G. Palla et al. "Uncovering the overlapping community structure of complex networks in nature and society". In: *Nature* 435.7043 (2005), pp. 814–818.

[27] J. Yang and J. Leskovec. "Overlapping community detection at scale: a non-negative matrix factorization approach". In: *Proceedings of the sixth ACM international conference on Web search and data mining*. 2013, pp. 587–596.

[28] A. F. McDaid, D. Greene, and N. Hurley. "Normalized Mutual Information to evaluate overlapping community finding algorithms". In: *New Journal of Physics* 11 (Mar. 2009). DOI: 10.1088/1367-2630/11/3/033015.

[29] T. O. Kvålseth. "On Normalized Mutual Information: Measure Derivations and Properties". In: *Entropy* 19.11 (Nov. 2017). DOI: 10.3390/e19110631.

[30] L. M. Collins and C. W. Dent. "Omega: A general formulation of the Rand index of cluster recovery suitable for non-disjoint solutions". In: *Multivariate Behavioral Research* 23.2 (Apr. 1988), pp. 231–242.

[31] V. da Fonseca Vieira, C. Ribeiro Xavier, and A. G. Evsukoff. "A comparative study of overlapping community detection methods from the perspective of the structural properties". In: *Applied Network Science* 5.1 (Aug. 2020), p. 51. DOI: 10.1007/s41109-020-00289-9.

[32] A. Lancichinetti et al. "Characterizing the community structure of complex networks". In: *PLoS ONE* 5.8 (2010). Available: http://dx.doi.org/10.1371/journal.pone.0011976.

[33] Z. Ghalmane et al. "Exploring Hubs and Overlapping Nodes Interactions in Modular Complex Networks". In: *IEEE Access* 8 (2020), pp. 79650–79683. DOI: 10.1109/ACCESS.2020.2991001.

[34] I. Sendiña-Nadal et al. "Assortativity and leadership emerge from anti-preferential attachment in heterogeneous networks". In: *Scientific Reports* 6.1 (Feb. 2016), p. 21297. ISSN: 2045-2322. DOI: 10.1038/srep21297.

[35] X. Zhang et al. "Logistic regression with network structure". In: *Statistica Sinica* 30.2 (Apr. 2020). Available: https://doi.org/10.5705/ss.202017.0281, pp. 673–693. DOI: 10.5705/ss.202017.0281.

[36] J. Leskovec and A. Krevl. *SNAP Datasets: Stanford Large Network Dataset Collection*. Available: http://snap.stanford.edu/data. June 2014.

[37] M. Zitnik et al. *BioSNAP Datasets: Stanford Biomedical Network Dataset Collection*. Available: http://snap.stanford.edu/biodata. Aug. 2018.

[38] Gabor Csardi and Tamas Nepusz. *The igraph software package for complex network research*. Available: https://igraph.org. 2006.

[39] B. Rozemberczki, O. Kiss, and R. Sarkar. "Little Ball of Fur: A Python Library for Graph Sampling". In: *Proceedings of the International Conference on Information and Knowledge Management*. 2020, pp. 3133–3140. DOI: 10.1145/3340531.3412758.

[40] C. L. Staudt, A. Sazonovs, and H. Meyerhenke. *NetworKit: A Tool Suite for Large-scale Complex Network Analysis*. Available: http://arxiv.org/abs/1403.3005. 2015.

[41] W. Xu et al. "A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models". In: *Signal Processing* 93.1 (Jan. 2013), pp. 261–276. DOI: 10.1016/j.sigpro.2012.08.005.

[42] C. Shao et al. "Rank correlation between centrality metrics in complex networks: An empirical study". In: *Open Physics* 16.1 (2018), pp. 1009–1023. DOI: 10.1515/phys-2018-0122.

[43] U. Brandes. "A Faster Algorithm for Betweenness Centrality". In: *Journal of Mathematical Sociology* 25.2 (2001), pp. 163–177. DOI: 10.1080/0022250X.2001.9990249.

[44] E. Bergamini and H. Meyerhenke. *Approximating Betweenness Centrality in Fully-dynamic Networks*. Available: http://arxiv.org/abs/1510.07971. 2015.

[45] J. Leskovec and C. Faloutsos. "Sampling from Large Graphs". In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*. Philadelphia, PA, USA, 2006, pp. 631–636. DOI: 10.1145/1150402.1150479.

[46] S. L. King. "Using ROC curves to compare neural networks and logistic regression for modeling individual noncatastrophic tree mortality". In: *Proceedings of the 13th Central Hardwood Forest Conference; Gen. Tech. Rep. NC-234*. Ed. by J. W. Van Sambeek et al. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station, 2003, pp. 349–358.

[47] C. Arimie, E. Biu, and M. Ijomah. "Outlier Detection and Effects on Modeling". In: *Open Access Library Journal* 7 (2020), pp. 1–30. DOI: 10.4236/oalib.1106619.

[48] J. Yang and J. Leskovec. "Community-Affiliation Graph Model for Overlapping Network Community Detection". In: *2012 IEEE 12th International Conference on Data Mining*. 2012, p. 1179. DOI: 10.1109/ICDM.2012.139.

[49] Rogier Noldus and Piet Van Mieghem. "Assortativity in complex networks". In: *Journal of Complex Networks* 3.4 (Mar. 2015), pp. 507–542. ISSN: 2051-1310. DOI: 10.1093/comnet/cnv005. eprint: https://academic.oup.com/comnet/article-pdf/3/4/507/2328341/cnv005.pdf. URL: https://doi.org/10.1093/comnet/cnv005.

APPENDICES

A Local Pluralistic Homophily for the text network

This appendix presents a table with the local pluralistic homophily and node properties in a text network, both with and without stopwords.

Table 3: Local Pluralistic Homophily and Node Properties in a Text Network.

Word	With Stopwords			No Stopwords		
	h_{word}	d_{word}	m_{word}	h_{word}	d_{word}	m_{word}
a	0.0000	4	2	0.0000	4	2
and	0.0000	13	2	0.0000	13	2
any	0.0114	2	1	0.0114	2	1
apoidea	-0.0114	4	1	-0.0052	4	1
both	0.0000	4	2	0.0000	4	2
by	0.0000	3	2	0.0000	3	2
characterize	0.0000	4	2	0.0065	4	2
chew	0.0000	4	2	0.0183	4	1
for	0.0000	4	1	0.0000	4	1
gather	0.0000	4	2	0.0418	4	1
hairybodied	0.0457	4	1	0.0418	4	1
honeybee	0.0000	4	2	0.0359	4	2
hymenopteran	0.0229	4	1	0.0418	4	1
include	0.0114	4	1	0.0065	4	2
insect	0.0000	4	2	0.0418	4	1
mouthpart	0.0000	4	2	0.0418	4	1
nectar	0.0000	4	2	0.0314	3	1
numerous	0.0343	4	1	0.0209	2	1
of	0.0000	7	2	0.0000	7	2
pollen	0.0000	2	1	0.0209	2	1
social	0.0000	3	2	0.0490	3	2
solitary	0.0000	4	2	0.0359	4	2
specie	0.0000	6	3	0.0817	5	2
sting	0.0343	4	1	0.0418	4	1
such	0.0000	4	2	-0.0052	4	1
suck	0.0000	4	2	0.0183	4	1
superfamily	0.0000	4	2	0.0418	4	1
the	-0.0229	8	4	-0.0229	8	4
usually	0.0457	4	1	0.0457	4	1
wing	0.0343	4	1	0.0314	3	1

Note: h_{word} is the pluralistic homophily of the node (word) in the network, d_{word} is the degree, and m_{word} is the number of memberships. Values colored indicate homophily classification: green for assortative and red for disassortative. Non-colored are non-assortative.

B Residual Analysis for Linear and Quadratic Models

This appendix shows the residuals analysis for both linear and quadratic models. The table below summarizes the mean and standard deviation of the residuals for each network and metric.

Table 4: Residual Analysis for Linear and Quadratic Models.

Network	Metric	Linear Mean	Linear Std	Quadratic Mean	Quadratic Std
StackOverflow	closeness	-1.845e-21	5.3e-05	1.107e-19	5.3e-05
StackOverflow	degree	-4.612e-22	5.3e-05	-9.778e-20	5.1e-05
StackOverflow	betweenness	-4.612e-22	5.3e-05	-9.225e-22	5.2e-05
StackOverflow	eigenvector	0.e+00	5.4e-05	2.767e-21	5.3e-05
DBLP	closeness	-7.21e-21	2.7e-05	-2.874e-18	2.5e-05
DBLP	degree	3.09e-21	2.7e-05	-1.324e-19	2.7e-05
DBLP	betweenness	1.03e-21	2.9e-05	1.03e-21	2.9e-05
DBLP	eigenvector	0.e+00	2.9e-05	2.06e-21	2.9e-05
Amazon	closeness	-1.156e-21	1.9e-05	-9.078e-20	1.9e-05
Amazon	degree	-1.156e-21	1.8e-05	-3.587e-19	1.8e-05
Amazon	betweenness	-2.120e-21	1.9e-05	-1.156e-21	1.9e-05
Amazon	eigenvector	-1.542e-21	1.9e-05	-7.71e-22	1.9e-05
LiveJournal	closeness	4.307e-22	1.9e-05	2.743e-19	1.9e-05
LiveJournal	degree	-6.46e-22	1.7e-05	1.077e-21	1.7e-05
LiveJournal	betweenness	-2.153e-22	1.8e-05	-2.153e-22	1.7e-05
LiveJournal	eigenvector	-4.307e-22	1.9e-05	1.723e-21	1.9e-05
YouTube	closeness	-2.455e-20	2.82e-04	6.248e-18	2.66e-04
YouTube	degree	-8.565e-21	1.6e-04	-2.227e-20	8.9e-05
YouTube	betweenness	-9.457e-22	1.22e-04	-2.712e-21	8.5e-05
YouTube	eigenvector	-1.142e-21	1.95e-04	-5.068e-21	1.23e-04
Orkut	closeness	5.220e-21	1.35e-04	8.018e-19	1.34e-04
Orkut	degree	-4.176e-21	1.23e-04	1.098e-18	8.3e-05
Orkut	betweenness	0.e+00	1.33e-04	8.353e-21	1.32e-04
Orkut	eigenvector	0.e+00	1.25e-04	-4.385e-20	8.e-05
Protein-Protein Interaction	closeness	-4.346e-20	6.e-05	2.320e-19	4.9e-05
Protein-Protein Interaction	degree	-6.518e-21	3.2e-05	1.308e-18	3.1e-05
Protein-Protein Interaction	betweenness	0.e+00	5.7e-05	-6.953e-21	5.7e-05
Protein-Protein Interaction	eigenvector	-1.738e-21	6.6e-05	1.512e-19	6.3e-05
Drug-Drug Interaction	closeness	-3.644e-21	1.48e-04	4.122e-18	1.45e-04
Drug-Drug Interaction	degree	0.e+00	1.47e-04	-3.644e-21	1.45e-04
Drug-Drug Interaction	betweenness	-7.289e-21	1.48e-04	3.644e-21	1.48e-04
Drug-Drug Interaction	eigenvector	7.289e-21	1.44e-04	-9.475e-20	1.41e-04
C. Elegans	closeness	-4.168e-19	3.931e-03	-6.399e-18	2.521e-03
C. Elegans	degree	-3.862e-19	2.289e-03	-3.578e-19	4.56e-04
C. Elegans	betweenness	6.13e-21	9.72e-04	6.13e-21	8.59e-04
C. Elegans	eigenvector	-2.452e-19	3.073e-03	4.766e-19	1.392e-03

C Detailed Coefficients of Logistic Regression Models Across Networks

This appendix presents a detailed breakdown of the coefficients derived from logistic regression models applied across various networks. Each entry corresponds to the influence of a specific centrality metric on the classification of nodes into homophily categories, measured across different networks. The table includes the network name, homophily category (Assortative, Non-Assortative, Disassortative), centrality measure used in the model (e.g., degree, closeness, eigenvector, betweenness), logistic regression coefficient indicating the influence of the centrality measure on the likelihood of a node belonging to a specific class, and the standard error of the coefficient (SE_Coefficient), providing a measure of the estimate's precision and variability.

Table 5: Logistic Regression Coefficients Across Networks

Network	**Class**	**Metric**	**Coefficient**	**SE_Coefficient**
StackOverflow	Assortative	degree	2.19366	0.0468415
StackOverflow	Assortative	closeness	2.06878	0.0213383
StackOverflow	Assortative	eigenvector	-1.24787	0.0372209
StackOverflow	Assortative	betweenness	0.246651	0.0402726

Continued on next page

Table 5 – continued from previous page

Network	Class	Metric	Coefficient	SE_Coefficient
StackOverflow	Non-Assortative	degree	2.39202	0.0468415
StackOverflow	Non-Assortative	closeness	0.871283	0.0213383
StackOverflow	Non-Assortative	eigenvector	-0.358882	0.0372209
StackOverflow	Non-Assortative	betweenness	0.241851	0.0402726
StackOverflow	Disassortative	degree	-4.58568	0.0468415
StackOverflow	Disassortative	closeness	-2.94007	0.0213383
StackOverflow	Disassortative	eigenvector	1.60675	0.0372209
StackOverflow	Disassortative	betweenness	-0.488502	0.0402726
DBLP	Assortative	degree	2.3711	0.0220743
DBLP	Assortative	closeness	0.282078	0.020322
DBLP	Assortative	eigenvector	-0.313902	0.0156878
DBLP	Assortative	betweenness	-0.170594	0.0154158
DBLP	Non-Assortative	degree	2.86857	0.0220743
DBLP	Non-Assortative	closeness	0.150786	0.020322
DBLP	Non-Assortative	eigenvector	-0.505698	0.0156878
DBLP	Non-Assortative	betweenness	-0.141186	0.0154158
DBLP	Disassortative	degree	-5.23967	0.0220743
DBLP	Disassortative	closeness	-0.432864	0.020322
DBLP	Disassortative	eigenvector	0.819599	0.0156878
DBLP	Disassortative	betweenness	0.31178	0.0154158
Amazon	Assortative	degree	1.13708	0.0146047
Amazon	Assortative	closeness	-0.066812	0.0132797
Amazon	Assortative	eigenvector	-0.182436	0.0126388
Amazon	Assortative	betweenness	-0.172817	0.0140922
Amazon	Non-Assortative	degree	-0.229324	0.0146047
Amazon	Non-Assortative	closeness	0.437363	0.0132797
Amazon	Non-Assortative	eigenvector	0.00902248	0.0126388
Amazon	Non-Assortative	betweenness	0.0830541	0.0140922
Amazon	Disassortative	degree	-0.907754	0.0146047
Amazon	Disassortative	closeness	-0.37055	0.0132797
Amazon	Disassortative	eigenvector	0.173413	0.0126388
Amazon	Disassortative	betweenness	0.0897631	0.0140922
LiveJournal	Assortative	degree	0.453394	0.0316807
LiveJournal	Assortative	closeness	2.43152	0.0238051
LiveJournal	Assortative	eigenvector	0.324673	0.0213971
LiveJournal	Assortative	betweenness	-0.0662377	0.0245862
LiveJournal	Non-Assortative	degree	0.420104	0.0316807
LiveJournal	Non-Assortative	closeness	-0.0216151	0.0238051
LiveJournal	Non-Assortative	eigenvector	0.158642	0.0213971
LiveJournal	Non-Assortative	betweenness	0.0226928	0.0245862
LiveJournal	Disassortative	degree	-0.873498	0.0316807
LiveJournal	Disassortative	closeness	-2.4099	0.0238051
LiveJournal	Disassortative	eigenvector	-0.483315	0.0213971
LiveJournal	Disassortative	betweenness	0.0435449	0.0245862
YouTube	Assortative	degree	-1.33008	0.069204
YouTube	Assortative	closeness	-0.215824	0.025807
YouTube	Assortative	eigenvector	-0.471056	0.0671987
YouTube	Assortative	betweenness	-0.131959	0.0305099
Orkut	Assortative	degree	1.50677	0.0781908
Orkut	Assortative	closeness	1.06722	0.0249934
Orkut	Assortative	eigenvector	-1.13266	0.0759665
Orkut	Assortative	betweenness	0.108048	0.0241184
Orkut	Non-Assortative	degree	1.43397	0.0781908
Orkut	Non-Assortative	closeness	-0.538836	0.0249934
Orkut	Non-Assortative	eigenvector	0.856251	0.0759665
Orkut	Non-Assortative	betweenness	-0.0931393	0.0241184
Orkut	Disassortative	degree	-2.94074	0.0781908
Orkut	Disassortative	closeness	-0.528388	0.0249934
Orkut	Disassortative	eigenvector	0.276408	0.0759665
Orkut	Disassortative	betweenness	-0.0149086	0.0241184
Protein-Protein Interaction	Assortative	degree	5.02043	0.0461459
Protein-Protein Interaction	Assortative	closeness	2.1888	0.026943
Protein-Protein Interaction	Assortative	eigenvector	-2.92145	0.0325195
Protein-Protein Interaction	Assortative	betweenness	0.369969	0.0275356
Protein-Protein Interaction	Non-Assortative	degree	-0.257625	0.0461459
Protein-Protein Interaction	Non-Assortative	closeness	-0.743376	0.026943
Protein-Protein Interaction	Non-Assortative	eigenvector	1.40066	0.0325195
Protein-Protein Interaction	Non-Assortative	betweenness	-0.864081	0.0275356
Protein-Protein Interaction	Disassortative	degree	-4.76281	0.0461459
Protein-Protein Interaction	Disassortative	closeness	-1.44542	0.026943
Protein-Protein Interaction	Disassortative	eigenvector	1.52079	0.0325195
Protein-Protein Interaction	Disassortative	betweenness	0.494111	0.0275356
Drug-Drug Interaction	Assortative	degree	-0.495996	0.145626
Drug-Drug Interaction	Assortative	closeness	0.480915	0.0678417
Drug-Drug Interaction	Assortative	eigenvector	0.954349	0.136631
Drug-Drug Interaction	Assortative	betweenness	-0.148832	0.0435268
Drug-Drug Interaction	Non-Assortative	degree	1.29049	0.145626
Drug-Drug Interaction	Non-Assortative	closeness	0.366068	0.0678417

Continued on next page

Table 5 – continued from previous page

Network	Class	Metric	Coefficient	SE_Coefficient
Drug-Drug Interaction	Non-Assortative	eigenvector	-1.22642	0.136631
Drug-Drug Interaction	Non-Assortative	betweenness	0.145243	0.0435268
Drug-Drug Interaction	Disassortative	degree	-0.794495	0.145626
Drug-Drug Interaction	Disassortative	closeness	-0.846983	0.0678417
Drug-Drug Interaction	Disassortative	eigenvector	0.272072	0.136631
Drug-Drug Interaction	Disassortative	betweenness	0.0035896	0.0435268
C. Elegans	Assortative	degree	-1.10935	0.436204
C. Elegans	Assortative	closeness	-0.259813	0.138752
C. Elegans	Assortative	eigenvector	-0.720739	0.393191
C. Elegans	Assortative	betweenness	0.0677464	0.186402

D Preliminary Work: Overlapping Community Detection in StackOverflow

The aim of the preliminary work is to study if the overlapping communities detected in StackOverflow network relate to node types. We use four representative algorithms to identify overlapping communities. We compare the results of algorithms with the composite performance framework defined in [12]. The best results are obtained using Hierarchical Link Clustering method (HLC, also referred to as links method). Also, we look for a hierarchical structure in the resulting overlapping communities. After determined the strength of the hierarchical structure of the overlapping communities, we test an initial metric for pluralistic homophily for various hierarchical levels within this structure. The metric is defined as a function of quantity of overlap of nodes between communities. Our preliminary results show that there exists assortative mixing in the network (homophily with respect to the proposed metric) and that assortative mixing depends on the level of the hierarchical structure. Broadly speaking, the results suggest that there is a significant relationship between overlapping communities and the types of nodes. Most importantly, such relationships can be studied with a pluralistic homophily approach.

D.1 Dataset Description

The raw dataset is the StackOverflow data contains the posts of questions and answers, as well as the tags associated with questions and information about which users participate in the discussion. Tags are an important part of the posts because they allow us to classify the topics of the questions. The data files include posts from 2008 to 2017, but we focus on the data starting from 2013. Table 6 shows the numbers for the general activity of StackOverflow reflected in active users, questions, answers and used tags for each year. An active user is one that makes questions, answers, comments or votes on a post.

Table 6: Users Activity per Year of the StackOverflow Site.

Feature	2013	2014	2015	2016	2017
Active Users	781,401	895,971	1,006,101	1,104,795	1,146,414
Number of Questions	2,052,675	2,155,813	2,306,151	2,477,353	2,376,885
Number of Answers	3,331,678	3,220,021	3,178,499	3,114,934	2,904,939
Number of Tags	6,080,455	6,468,815	6,937,356	7,415,487	7,415,487
Number of Unique Tags	32,182	34,992	37,891	40,984	42,442

The StackOverflow activity reflects a continuous increase in questions made from 2013 to 2016 and a small decreasing in 2017. The number of answers for those questions has been more or less stable with a low tendency to decrease. This behavior may be explained because the more questions are answered, the fewer questions are made about a specific topic. With regard to the tags attached to the questions, it shows that the number of tags every year increases in the same proportion as the number of questions. Also, consistent the years, unique tags increase almost in the same proportion (note that some tags are more frequent than others). However, as a question can be related to several topics at the same time, an important factor that needs to be considered is the co-occurrence of tags within the same question. At most five tags are permitted in every question. Figure 10 shows the most frequent co-occurrence of tags in questions for each year. These co-occurrences suggest the existence of hierarchical relation between technologies represented by tags in the same question, where a specific technology could be a specialization of another. Note that a relationship may also indicate the joint usage of two or more technologies. For instance, a co-occurrence between *<javascript>* and *<jquery>* could be interpreted as a relation of hierarchy, while a co-occurrence of *<java>* and *<mysql>* could be a relation of joint use.

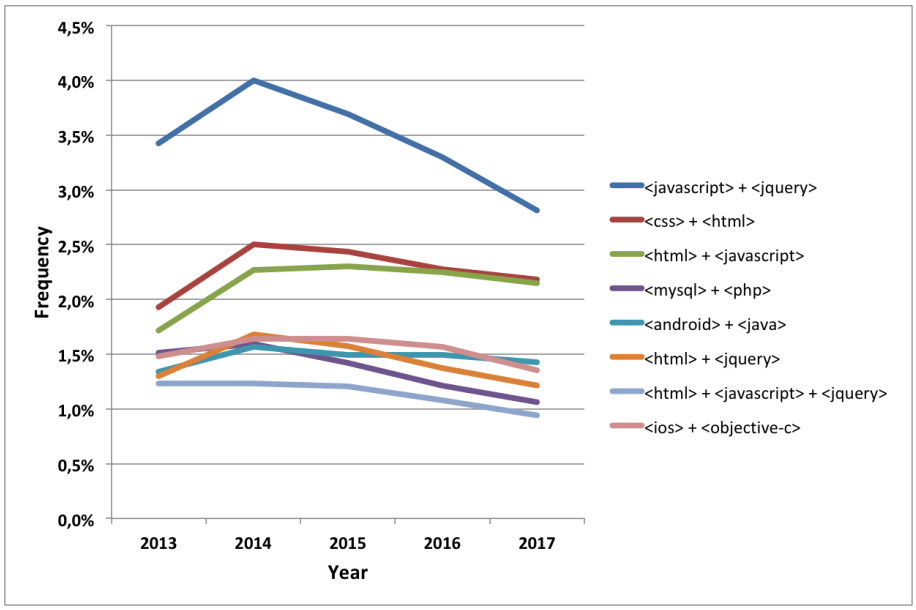


Figure 10: Top most co-occurrence tags frequency by year.

Based on these preliminary results, we suggest the possibility to find communities of users around technologies. Furthermore, if such communities have a hierarchical relationship, this is probably related to the co-occurrence of tags in the questions, which can be interpreted as the formation of clusters of technologies. This gives us a direction of how to build a network model that represents the programmer's communities in StackOverflow in such a way that they can belong to multiple ones. With regard to which communities a programmer belongs to, we suggest that they are those that have a strong relationship with the technologies of their interest. In other words, our network model represents overlapping communities of programmers on StackOverflow.

D.2 Empirical network model for StackOverflow

Based on previous data characterization, we propose to construct a graph that consists of nodes that represent users and edges that represent affiliations between them. An affiliation between two users means that one user has answered a question from the other one. In our analysis, we consider the network mainly as undirected. In this network, it does not matter which of the two users makes the question or which answers it. We also consider, only for purposes of comparison of the performance of community detection algorithms, a directed network. In this case, the edge goes from the user who asks the user who responds. For both kinds of networks, the weight of an edge represents the final score of the votes given by the users to the answer. This network will henceforth be referred to as the *network of users*. Figure 11 depicts its basic network structure.

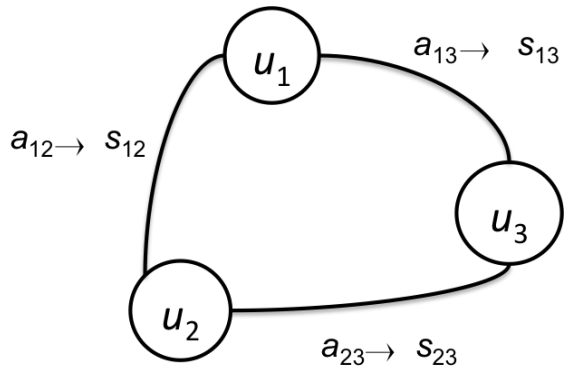


Figure 11: Basic graph structure representing Network of Users.

Formally, let $U = \{u_1, u_2, \ldots, u_n\}$ the set of active users, $A = \{a_{ij} \mid a_{ij} = \{u_i, u_j\} : u_i, u_j \in U\}$ the set of answers of user u_i to questions from user u_j, and $s : a_{ij} \rightarrow s_{ij}$ is the score given to each answer, then a network of users in StackOverflow is defined as a triple $N_u = (U, A, s)$.

We study basic metrics represented in density, mean degree and assortativity. For simplicity, we considered a sub-graph representing the largest connected component of the entire network. Table 7 shows the measures for the network over the years 2013-2017.

Table 7: Basic Measures of Network of Users

Feature	2013	2014	2015	2016	2017
Number of Nodes	55,493	44,772	37,294	29,648	19,640
Number of Edges	367,724	228,288	161,524	112,319	63,630
Density	0.00011941	0.00011388	0.000116134	0.00178739	0.00016497
Mean degree	13.25	10.20	8.66	7.57	6.48
Assortativity	-0.00321512	-0.01242976	-0.01345701	-0.01334031	-0.0061545

Note that in the network of users both the number of edges and the degree of the network decreases every year. This is congruent with the observed in table 6 where answers decrease every year. The fewer responses fewer relationships between users can be established.

D.3 Overlapping Community Detection

There are a large number of detection algorithms for detecting overlapping communities [16]. We ran Hierarchical Link Clustering (HLC) [12], Clique Percolation [13], Greedy Modularity[14] and Infomap [15] over data extracted for 2017 year.

Table 8 shows a comparison of basic numbers resulting from the overlapping community detection algorithms.

Table 8: Communities Detected by Overlapping Community Algorithms

Algorithm	Number of Communities	Size of largest Community (members)	Mean of Community Size
Clique Percolation	5,415	13,317	5.9
Greedy Modularity	3,768	121,346	132.3
Links	85,720	4,451	877.0
Infomap	36,559	3,541	136.4

HLC yields the largest number of communities detected with the largest average community size. Unlike traditional community detection methods based on comparing the similarity of the nodes, HLC makes a comparison of similarity between links. Links are arranged in a dendrogram based on their similarity that is calculated with a *Jaccard index* (or *Tanimoto index* for weighted similitudes). The dendrogram is cut at different thresholds (from bottom to up) until average density for all resulting partitions of links gives the maximum value. The communities are formed, one for every given partition of links. As each partition of links has correspondence with a single community, the memberships of the nodes to the communities are assigned depending on what nodes participate in each link. In this way, a node can be overlapped in as many communities as it has links in different partitions.

However, table 8 not reflect the quality nor the coverage of overlapping. Such measures are described in the following section.

D.4 Performance Evaluation of Methods

Although there is not a widely accepted framework for comparing the performance of community detection algorithms, the quality and quantity of communities are common comparison metrics. The proposed approach borrows ideas from the performance comparison framework defined in [12]. This framework allows for fair comparisons between detection algorithms. In general, certain methods perform well finding quality communities but do not cover the entire network or vice versa. We also note that there are two dimensions in the framework. The first framework defines the aspects to be measured: communities and overlapping; the second framework defines the metrics: quality and coverage. With this approach, it is possible to have four metrics to resolve, in a quantitative way, the questions overviewed in Figure 12.

		Aspect	
		Community	Overlap
Metric	Quality	How homogeneous is each community?	How accurate is the # of overlap?
	Coverage	How many nodes are covered?	How many memberships are assigned?

Figure 12: Composite Framework Structure.

The four metrics that compose the performance framework are *community quality*, *overlap quality*, *community coverage*, and *overlap coverage*. The first two metrics are based on metadata and the last two metrics focus on the amount of information extracted from the network.

Metadata is an important element in the framework so, although does not directly used in the construction of the network, it gives information about the types of certain nodes. In the case of the StackOverflow network of users, the metadata represents the tags attached to every question that reflects, by transitivity, which technologies are related to the users that made or answers posts. In other words, the metadata for each user is the set of tags that have been related to questions and answers of the user over time.

Below we give a brief definition and explanation of each metric:

1. Community quality: The measure computes, based on metadata, how the degree of similarity between nodes that belong to the same community. It is defined through the following equation

$$E = \frac{\langle \mu(i,j) \rangle_{\text{ all i,j within same community}}}{\langle \mu(i,j) \rangle_{\text{ all pairs i,j}}} \tag{24}$$

E defines the *enrichment* of node pair similarity, where $\mu(i,j)$ is a metadata-based similarity between node i and j. In consequence, in the network of users, enrichment is defined as the average similarity between users that share the same community, divided by the average similarity of all pairs of users in the network, where both numerator and denominator, are calculated based on metadata that gives information about what tags are related with each user. To measure the similarity between a pair of users, we calculate the Jaccard index between their tags sets as follow

$$\mu(i,j) = \frac{|T_i \cap T_j|}{|T_i \cup T_j|} \tag{25}$$

where T_u is the set of tags associated with user u. With this similarity, the community quality is then calculated using equation 24.

It is expected that a good algorithm detects communities of users where there are more tags in common between users of same communities, which means communities with more similar users and therefore the more enrichment within communities.

2. Overlap quality: The measure uses *mutual information* to relate, for each node of the network, the number of memberships assigned with the number of true communities that have been extracted from the overlap metadata. In the network of users, overlap metadata are the multiples tags associated with a user, where each tag represents a technological community that it belongs to. The measure of mutual information I for the networks of users is defined as

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \ log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{26}$$

where X is the number of communities for each user u, Y is the number of tags for each user u in the overlap metadata, $p(x,y)$ is the joint probability function of x and y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y respectively.

It is expected that the best algorithm detects communities maximizing quantity of mutual information, which tell us that users with more tags belong to more communities (due to the broader interest of such users for multiple technologies).

3. Community coverage: This measure represents the fraction of nodes that belong to at least one nontrivial community, defined as a community that is composed of three or more users. The community coverage H is defined as

$$H = \left(\frac{|\bigcup c_i|}{|S|} \right) \tag{27}$$

where $c_i \subseteq C$ and C is the family of detected communities of users, such that $|c_i| > 2$

A good measure of community coverage indicates that the largest amount of information on the network was analyzed.

4. Overlap coverage: This measure indicates how many densely overlapping communities were extracted by the algorithm. It counts the average number of memberships of the users who belong to nontrivial communities. Overlap Coverage O is defined as

$$O = \frac{\langle \sum ms(m_i) \rangle}{|S|} \tag{28}$$

where $m_i \in M = \bigcup c_i \subseteq C$ and C is the family of detected communities of users, such that $|c_i| > 2$, and ms is a function that counts the number of memberships of m.

In some cases, these measures do not necessarily fall between 0 and 1 so the results are normalized so that the maximum value of 1 represents the best performing algorithm. Summing all four metrics the maximum composite performance is 4. We compute and compare, for all algorithms, the measures of the composite performance indicated above. We considered first an undirected network of users, that is, a network where an undirected link between two users represents an interaction, no matters who asked or who answered a post. Also, we considered two possible variants of the undirected network, one weighted (fig. 13) where the sum of score voting gives the weight of the link and the other an unweighted with no scores given (fig. 14). For both variants, the results show clearly a better performance for HLC followed by Infomap. The algorithms with a result of 1 in a measure have the best performance. Although HLC only outperforms in one measure the other algorithms, it has the best composite performance. Note that the major difference between HLC and the other algorithms (except by Clique method) lies in the performance of overlap quality. However, overlap quality represents the worst performance in all other measures.

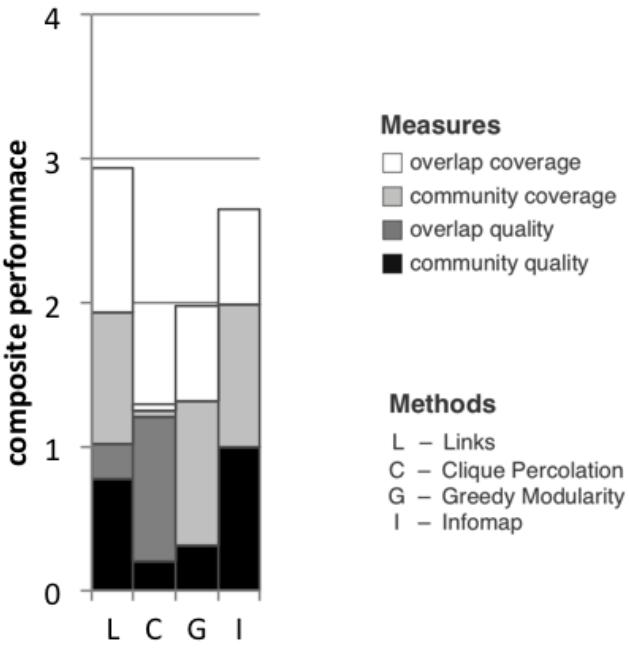


Figure 13: Weighted Network of Users.

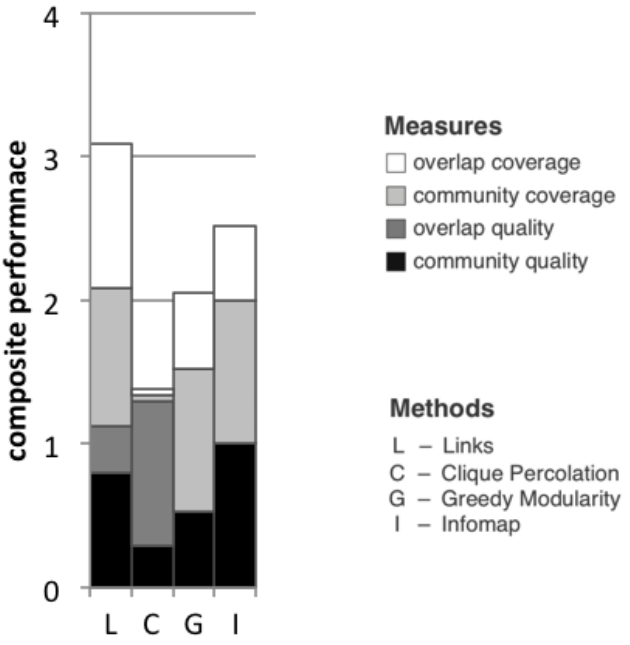


Figure 14: Unweighted Network of Users.

As shown in figures 13 and 14, for the unweighted network, the HLC algorithm has a better performance than for the weighted network. As the objective of the first comparison is to measure each one algorithm against each other, we also test the two variants of weighted and unweighted networks, obtaining that the weighted network has a slight advantage over the unweighted. Figure 15 shows this comparison.

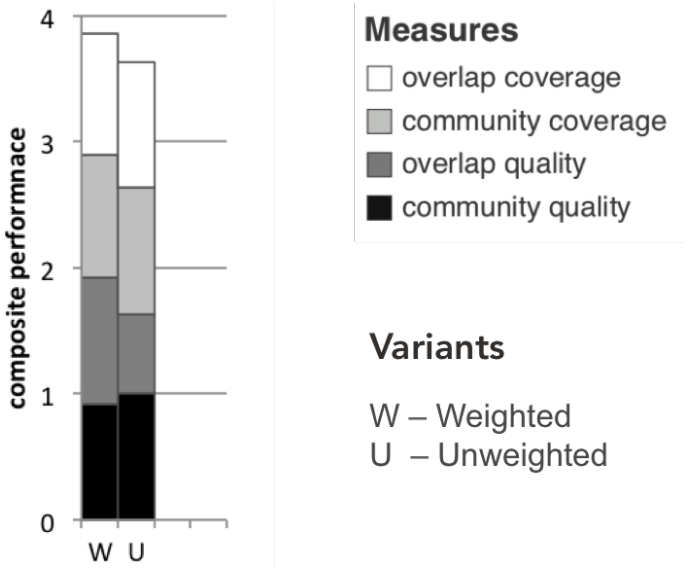


Figure 15: Weighted and Unweighted Network of User results comparison.

With the weighted network delivering the best performance we also compute the measures for the directed network. In this case, the link direction goes from the user who answers the question to the user posts the question. Since some algorithms cannot be applied to directed networks, we were only able to test HLC on the undirected network, HLC on the directed network and Infomap. The results were similar that the undirected network. The HLC on the undirected algorithm has the best performance, as shown in figure 16.

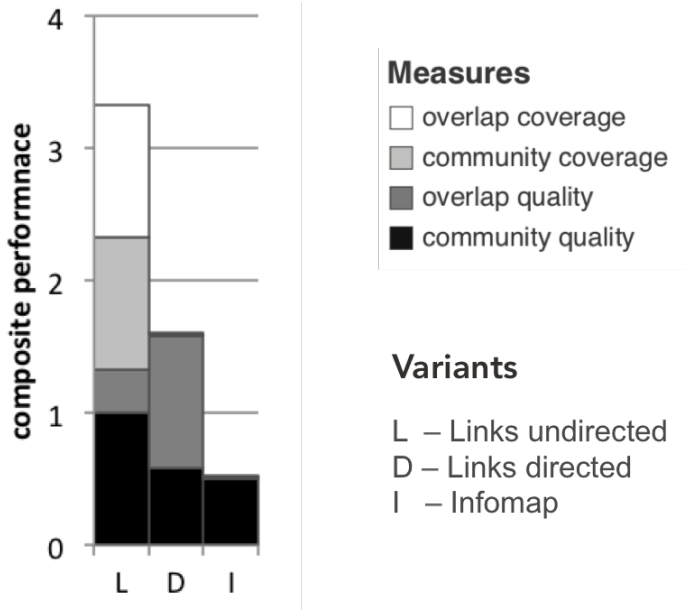


Figure 16: Directed Network of User results comparison.

Overall, HLC is the method with the best results in all possible scenarios. Thus, we will use HLC to detect communities on the undirected but weighted network

of users. This network serves as our baseline for the identification of hierarchical structures and measurement of pluralistic homophily.

D.5 Results and Discussion

Our preliminary results show a strong relationship between the overlapping communities and the types of nodes within them. We find evidence of different and meaningful hierarchical structures of overlapping communities detected in the StackOverflow network. Moreover, we show that every community present in a different level of the hierarchy is meaningful in the sense of represents in good shape specific technologies used by the members of such community. We also prove that the hierarchical structure can change significantly and that there is not a fixed community structure with overlapping communities. Overall, we find a strong relationship between the phenomenon of pluralistic homophily and the hierarchical structure of the detected overlapping communities.

First, to identify the hierarchical structure of the overlapping communities, we ran the HLC algorithm cutting the dendrogram at different points, ranging the threshold t from 0 to 1 in steps of $\Delta t = 0.1$. With the results of the number of communities detected for each threshold, we calculate the *branching probability*, that is, the fraction of communities that split into more communities at the threshold $t + \Delta t$ (see figure 17).

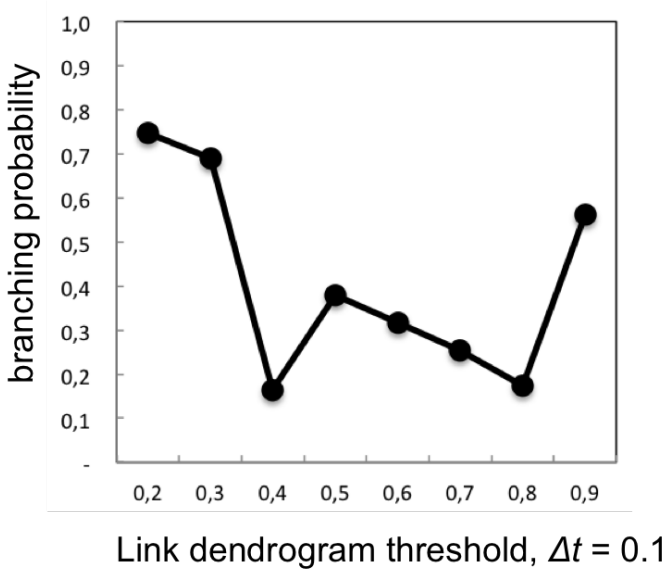
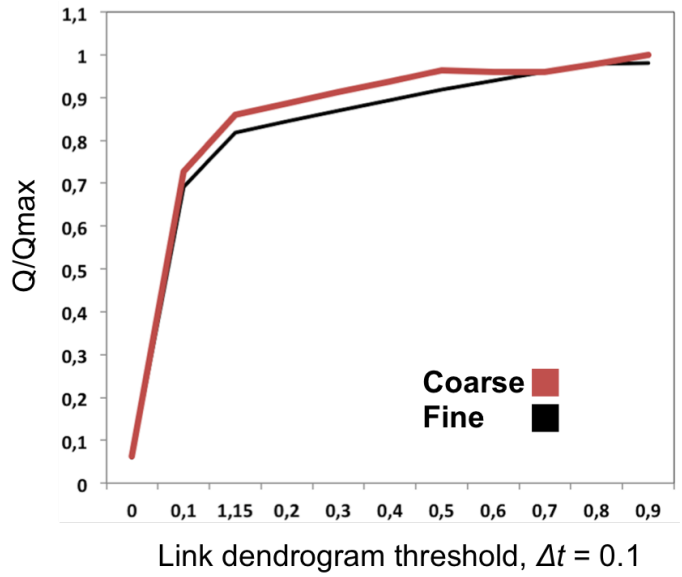


Figure 17: Branching probability of communities structure.

To determinate if the communities detected at various thresholds are meaningful, we use the tags in the metadata. As defined in [12], the metadata enable us to calculate the community quality of detected communities for each threshold. Since



Hierarchical Fine and Coarse metadata. $Q/Qmax$ is a normalized measure of community quality enrichment. Note that the threshold at 1.15 is an inflection point for both lines. Such threshold match with the maxim average density used by HLC to cut the dendrogram at the optimal point.

high thresholds detect communities more detailed than low ones, it is necessary to differentiate two levels of metadata: One *coarse* and one *fine*. The coarse metadata is composed by the most co-occurring tags as are shown in figure 10. Fine metadata are all remaining tags. The results in figure D.5 shown that that big sized communities detected at low thresholds conform with the coarse metadata while the small-sized communities at higher thresholds conforms with the fine metadata.

To have a first approximation towards a metric for pluralistic homophily we use the concept introduced in [48] which states that the similarity between two nodes is determinate by multiple dimensions besides node degree. The work in [27] defines pluralistic homophily as a similarity of nodes that is proportional to the number of shared memberships. Based on the definition of homophily with respect to a numeric property of the nodes [1], we define the pluralistic homophily as a function of the quantity of overlapping of the nodes in the network. Our first approach for the metric of pluralistic homophily, denoted by p, is

$$p = \frac{\sum_{ij}(A_{ij} - k_i k_j/2m)o_i o_j}{\sum_{ij}(k_i\delta_{ij} - k_i k_j/2m)o_i o_j} \tag{29}$$

where A is the adjacency matrix of nodes, k is the node degree, m is the number of edges in the network and o is the overlap quantity in a node. δ_{ij} is the Kronecker delta function (which equals 1, when both arguments are the same and 0, otherwise).

We apply equation 29 to the StackOverflow network, taking the memberships assigned by HLC in an optimal threshold, yielding $p = 0.05$. As a comparative,

we also calculate the assortativity degree of the network, denoted by r, obtaining a value of $r = -0.05$.

Notice that due to assortativity ranges goes from -1 to 1, which means that in the first instance that network is disassortative with respect to node degree, but assortative with respect to node overlap quantity. We hypothesize that users with low activity (users that ask or respond a few posts) tend to link with users with high activity (users that ask or respond many posts). However, users also tend to link with users that participate in a number of similar communities.

We also demonstrate a proof-of-concept of the relationship between the pluralistic homophily and the structure of the overlapping communities. Figure 18 shows the variation of p for the different groups of communities. This result shows that, from the perspective of pluralistic homophily, the tendency of nodes to link to other nodes, varies according to the different structures resulting from the hierarchy of communities.

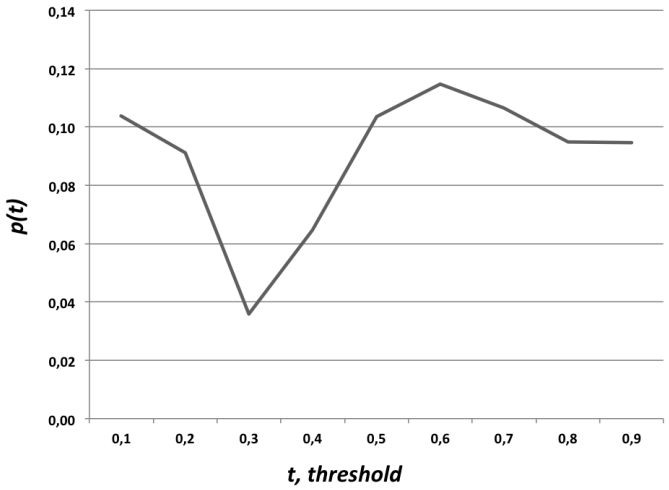


Figure 18: Pluralistic homophily according to structure of communities detected.

Our findings support the hypothesis that a node tends to link to other nodes of the same type, located with the same overlapping. Also, we find that this tendency varies according to the hierarchical structure of the communities detected in the network. We suggest that pluralistic homophily can measure the tendency of a node to link to other nodes based on the structure of overlapping communities; therefore, pluralistic homophily is a useful concept to estimate the type of a node. In the development of proposed research, we expect to analyze other approaches to measure pluralistic homophily, as for instance some based on local assortativity [49]. Likewise, we expect to validate other metrics, not only based on the quantity of node overlap, but also on the quality of overlapping, varying then the quantity of the metadata associated with the nodes of overlapping communities.