Definition of status and value assortativity in complex networks and their evaluation in Threads

(Article begins on next page)

Definition of status and value assortativity in complex networks and their evaluation in Threads

Gianluca Bonifazi[1], Enrico Corradini[1*], and Domenico Ursino[1]

[1] DII, Polytechnic University of Marche
[*] Contact Author
g.bonifazi@univpm.it; e.corradini@univpm.it; d.ursino@univpm.it

Abstract

The concept of "assortativity" was introduced to denote the preference of a network node to relate to other nodes similar in some way. This concept is closely related to that of homophily, which is used in Social Network Analysis to indicate that users tend to connect with others who have similar characteristics. It is possible to distinguish between two types of homophily, namely status homophily and value homophily. We believe that this distinction can be extended to assortativity. Therefore, in this paper we propose a novel conceptualization of assortativity by introducing status and value assortativity. This dual categorization provides a comprehensive understanding of both structural and content-based dimensions of network interactions. We apply these new assortativity concepts to Threads, a new social platform launched by Meta, and demonstrate their relevance and versatility in the analysis of emerging online platforms. Assortativity has been extensively studied in major social networks and various complex networks, but not in Threads, as this social platform is new. Introducing two types of assortativity and computing them for Threads are two important contributions of this paper. Actually, we provide a third. In fact, in the past literature, we did not find any dataset on Threads that could enable our assortativity analysis. To address this issue, we built a dataset and make it available to the scientific community for future investigations and analyses on Threads.

Keywords: Assortativity, Threads, Social Network Analysis, Status Homophily, Value Homophily

1 Introduction

The study of complex networks has significantly advanced our understanding of different aspects of complex systems, from interactions within biological ecosystems to the dynamics of social structures. One of the areas in which complex network analysis has produced the most significant results is Social Network Analysis. This has become a central research area with the advent of Online Social Networks (OSNs) and the need to discover patterns characterizing human interactions in OSNs and mechanisms regulating information flow in them.

Assortativity is a well-known concept in complex network analysis. It was introduced by Newman [19] with the goal of quantifying the propensity of nodes in a network to connect to other nodes that are somewhat similar. Similarity can cover a variety of aspects, such as structural aspects or characteristics of the objects represented by the network nodes. In the past, many authors have focused on structural aspects, in particular the degree of nodes, in which case we refer to degree assortativity. The assortativity, and consequently the disassortativity, of nodes can make an important contribution to determining the structural aspects of networks. They influence the coherence of the networks, the resilience to perturbations, and the efficiency of the processes to be analyzed; the latter can be related to the diffusion of information, the spread of epidemics, and the control of viruses [2, 30, 12, 1, 19, 26, 14, 3]. Newman's studies of assortativity have contributed greatly to our understanding of how complex networks work and evolve. These studies showed that social networks are generally assortative, with individuals tending to interact with others who have similar characteristics. They also showed that this tendency does not hold in other types of complex networks. For example, Newman showed that biological and technological networks are disassortative. These large differences in assortativity are indicative of the different ways in which different types of complex networks organize and evolve [20, 21].

In Social Network Analysis, assortativity is considered a special case of homophily. This concept, introduced by Lazarsfeld and Merton [17, 18], refers to a generic community and indicates the propensity of its members to interact primarily with other members with the same or otherwise similar characteristics [10, 6]. Sociologists have identified two types of homophily, namely status homophily and value homophily. The former refers to the tendency of individuals to interact with others of the same status (e.g., aristocrats with aristocrats, influencers with influencers, etc.). The latter refers to the tendency of individuals to interact with others who share the same values (e.g., they both care about climate change) [28]. Applying this concept to Social Network Analysis, status homophily refers to the network structure and represents, for example, situations in which high-degree nodes tend to interact with other high-degree nodes. In contrast, value homophily refers to the content exchanged and published by the users represented by the nodes in the network; in fact, the content reveals the reference values of the authors who publish it.

Considering the close relationship between assortativity and homophily in Social Network Analysis, we propose a new conceptualization of assortativity by introducing two innovative categories of it, namely status assortativity and value assortativity. Status assortativity focuses on the structural properties of nodes, while value assortativity considers the thematic content of user interactions. This dual perspective provides a more granular and comprehensive understanding of network dynamics, offering novel insights into both the structure and content of social networks. Consistent with what we have said about homophily in Social Network Analysis, status assortativity involves evaluating nodes from a structural point of view (e.g., based on their degree, betweenness or eigenvector centrality). In contrast, value assortativity involves evaluating nodes based on the content posted by the corresponding users. Thus, status assortativity indicates the preference of a node in a social network to correlate with other structurally similar nodes. Value assortativity denotes the preference of a node in a social network to correlate with other nodes in such a way that the corresponding users post similar content and thus have similar behaviors or mindsets. The introduction of the two categories of status assortativity and value assortativity is the first contribution of this paper.

Having introduced the new conceptualization of assortativity and the consequent categories of status assortativity and value assortativity, we decided to evaluate these ideas on Threads by computing appropriate forms of assortativity that can be traced back to these two categories. Threads is a new content-based social platform launched by Meta on July 6, 2023. Many experts see it as the natural competitor to X. Threads users can share text updates and initiate public conversations. Thanks in part to its ability to leverage Instagram's vast network of users, Threads recently reached more than 130 million active users. As a very new platform, Threads has not yet been extensively studied by social network researchers. In particular, to the best of our knowledge, no one has studied assortativity in Threads, whereas, as we will see below, there are many studies on this topic for older OSNs. The second contribution of this paper is precisely to study some forms of status assortativity and value assortativity on Threads.

In order to perform our assortativity analysis in Threads, we needed a dataset derived from this platform that contained all the structural and content information needed for this analysis. Unfortunately, we could not find an existing dataset in the literature that would support this type of analysis. Therefore, we decided to create one. After completing such a challenging task, we decided to make such a dataset open and available to all researchers who want to study and analyze Threads. And this is precisely the dataset made available to the scientific community that is the third contribution of this paper.

After constructing the Threads dataset, we made our assessments of status assortativity and value assortativity on it and compared our results to those already found in the literature for other complex networks, highlighting similarities and differences.

The structure of this paper is as follows: In Section 2, we present the related literature. In Section 3, we introduce the categories of status assortativity and value assortativity and propose some specific forms of them. In Section 4, we describe the main features of our Threads dataset. In Section 5, we illustrate our experiments. Finally, in Section 6 we draw our conclusions and highlight some possible developments of our research efforts.

2 Related Literature

The concept of assortativity was introduced by Newman in his famous paper [19]. In this paper, in addition to introducing the concept, he proposes a model for an assortative network and uses it for analytical and numerical studies. Through such studies he shows that assortative networks tend to percolate more easily than disassortative networks and are more robust to node removal. Finally, he analyzes different types of complex networks in terms of assortativity and shows that social networks are often assortative while technological and biological networks tend to be disassortative.

Since Newman's paper, many researchers have studied the assortativity of different types of networks, and OSNs have played a prominent role in these analyses.

For example, in [22] the authors provide a comprehensive overview of assortativity in complex networks, starting with Newman's original definition. In particular, they extend the original discussion focused on degree assortativity to include direct and weighted links. In [16], the authors explore the reasons why most networks, with the exception of social networks, exhibit disassortative mixing. Using an information theory approach, they show that some degree of disassortativity can be expected for scale-free networks because of the larger number of possible disassortative configurations. One of the first studies of assortativity in social networks is proposed in [11]. In this paper, the authors present a model of network growth tailored to describe social networks using the ArXiv:condmat preprint archive as a case study. In [15], the authors study degree correlation patterns of OSNs. They start from the common belief at the time that real social networks are assortative while OSNs are not. Through their studies, they show that OSNs exhibit a wide range of degree correlation patterns including disassortative, assortative, and nearly neutral mixing.

In [31], the authors analyze the evolution of network assortativity and show that it typically arises at the beginning of the network, then declines and finally stabilizes. Using empirical data from nine different social networks and a bidirectional selection model influenced by the Pareto distribution of wealth, they find a consistent pattern across networks. In [4], the authors explore the structure and dynamics of users in X. In particular, they introduce a method for measuring happiness in user-to-user communication and investigate the correlation of happiness levels across the network. In [5], the authors study the assortativity of Subjective Well-Being (SWB) in X. SWB can be thought of as a form of happiness. The authors find that X exhibits significant assortative mixing. In the same perspective, but for a completely opposite topic, the authors of [13] analyze the possible existence of a form of suicide-related assortativity in X, which they call Suicide-Related Verbalizations (SRVs). At the end of their analysis, they find that X is strongly assortative with respect to SRVs such that it affects not only immediate neighbors but also the second degree of separation, even when controlling for mood.

In [12], the authors propose an analysis of Reddit that examines subreddit and author behavior in order to identify several user stereotypes. They then analyze the possible existence of assortativity among users on the platform. At the end of this analysis, they find that Reddit is assortative with respect to three types of assortativity, the first based on degree centrality, the second based on eigenvector centrality, and the third based on users' co-posting activity. In [29], the authors propose a comprehensive structural analysis of the Facebook social graph. In particular, they examine the degree distribution, confirm the "six degrees of separation" phenomenon on a global scale, and show that the Facebook network is almost completely connected, with a single large component comprising 99.91% of users. The authors of [7] introduce a new form of assortativity in OSNs that focuses on membership overlap between different platforms. They also show that membership overlap is loosely assortative. This finding suggests that the phenomenon of users belonging to multiple online communities has a structured pattern that may be influenced by users' personality and social behavior. In [9], the authors introduce a form of assortativity that takes into account the tendency of users of one platform to be connected to other users with whom they share membership in other common platforms. The authors test this form of assortativity on Facebook and X and obtain a high value of the corresponding coefficient.

Assortativity has been much analyzed in X. In [27], the authors study a form of connectivity-based assortativity in this social platform. With respect to it, they find that X is very assortative. In fact, the value of the corresponding coefficient is very high. In [25], the authors study a form of assortativity on X based on shared interests and language. In [8], the authors define interest assortativity, which is a form of assortativity based on users' interests.

As stated in the Introduction, our paper stands out in the related literature by making some specific contributions. First, it introduces a classification of assortativity by deriving it from that of homophily. Indeed, similar to homophily, it distinguishes between status assortativity and value assortativity. Moreover, to the best of our knowledge, it is the first paper that studies assortativity in Threads. This is justified by the fact that the latter social network is very new. Finally, as a third contribution, it provides an anonymized Threads dataset that can be used by all researchers who want to study this social network.

3 Definition of status and value assortativity

In this section, we address the first goal of this paper, which is to propose a new conceptualization of the concept of assortativity by introducing two different categories of it, namely status assortativity and value assortativity. As mentioned in the Introduction, we take into account the close link between assortativity and homophily in Social Network Analysis and the distinction between status homophily and value homophily already known in the literature [28]. In particular, we provide the following definitions:

- *Status assortativity* considers structural properties of nodes. It indicates the tendency of a node in a network to correlate with other nodes that are structurally similar. In other words, it indicates the propensity of a node in a network to have arcs with nodes that have similar structural characteristics. For example, if we consider degree centrality as the reference structural characteristic, then status assortativity indicates the tendency of nodes with high (resp., low) degree centrality to connect with other nodes with high (resp., low) degree centrality rather than with nodes with medium or low (resp., high) degree centrality.
- *Value assortativity* focuses on the content and values reflected in user interactions. It denotes the tendency of a node in a network to correlate with other nodes such that the corresponding users show similar behaviors or mindsets (for instance, they publish similar content). In other words, it indicates the tendency of a node representing a user in a community to have arcs with nodes associated with users who share the same values. For example, if we take interest in a topic (e.g., climate change) and attitude toward it as reference characteristics, value assortativity reflects the tendency of nodes representing users who believe (resp., deny) climate change to be connected to other nodes associated with users holding the same belief.

Starting from these two definitions, in this section we propose two forms of status assortativity and one form of value assortativity that can be applied to any content-based social platform.

To do this, we must first define a model for representing a content-based network. Our model represents such a network as a directed graph:

$$\mathcal{N} = \langle N, A \rangle \tag{3.1}$$

N is the set of nodes in $\mathcal{N}$. A node $n_i \in N$ represents a user u_i of the content-based network. Since there is a biunivocal correspondence between a node and a user, we will use these two terms interchangeably in the following. Each node n_i has associated a label f_i indicating the number of followers of u_i.

A is the set of arcs of $\mathcal{N}$. An arc $a_{ij} = (n_i, n_j) \in A$ represents the set of interactions from a user u_i to a user u_j. An interaction from u_i to u_j indicates that u_i commented on a post published by

u_j. Each arc a_{ij} has an associated label TS_{ij}; the latter indicates the main topics discussed in the interactions from u_i to u_j and thus is an indicator of the content exchanged between them. Examples of possible topics include "Health", "Politics", "Technology", "Entertainment", etc.

Having introduced the model, we can define some forms of status assortativity and value assortativity for a content-based network. The first form of assortativity we consider is the most classical one, namely degree assortativity, which is clearly a form of status assortativity. The degree assortativity coefficient r_D corresponding to it can be defined as:

$$r_D = \frac{\sum_{a_{ij} \in A}(k_i - \overline{k})(k_j - \overline{k})}{\sum_{a_{ij} \in A}(k_i - \overline{k})^2} \tag{3.2}$$

Here, A is the set of arcs of $\mathcal{N}$, k_i (resp., k_j) is the degree (meant as the sum of the indegree and outdegree) of n_i (resp., n_j) and $\overline{k}$ is the average degree of the nodes of $\mathcal{N}$. In this formula, the numerator is obtained by computing the sum of the products of the degree deviations from the mean for each pair of connected nodes. It represents the covariance of the degrees between all pairs of connected nodes. A positive numerator indicates a tendency for high-degree nodes to connect with other high-degree nodes (and thus a tendency toward assortativity); on the other hand, a negative numerator indicates a tendency for high-degree nodes to connect with low-degree nodes (and thus a tendency toward disassortativity). The denominator represents the variance of the node degrees and acts as a normalizing factor. In fact, the values of r_D are in the real range $[-1, 1]$, where 1 represents perfect assortativity, -1 represents perfect disassortativity, and 0 represents the absence of any assortativity relationship.

Equation 3.2 aims to formalize the intuitive concept of status assortativity defined above. In fact, assuming that the structural characteristic of interest is degree centrality, the numerator considers each pair of nodes connected by an arc and provides a measure of how similar the nodes of the pair are in terms of their degree. To do this, it considers the difference between their degree and the average degree of all nodes. If both nodes in an arc have a degree greater than the average degree, it means that they tend to be similar, and their similarity grows the more they both have a degree much greater than the average degree. An analogous conclusion can be drawn if both nodes in an arc have a degree less than the average degree. On the other hand, if one node of an arc has a degree greater than the average degree and the other node has a degree less than the average degree, this means that they tend to be dissimilar, and their dissimilarity increases the more they both deviate from the average. The variance in the denominator serves only as a normalizing factor.

We have seen that our model associates each node n_i with the number of followers of the corresponding user u_i. Consequently, we can define a second form of status assortativity, which we call weighted degree assortativity. The weighted degree assortativity coefficient r_{WD} corresponding to it can be defined as follows:

$$r_{WD} = \frac{\sum_{a_{ij} \in A} \alpha(n_i, n_j) \cdot (k_i - \overline{k})(k_j - \overline{k})}{\sum_{a_{ij} \in A} \alpha(n_i, n_j) \cdot \frac{(k_i - \overline{k})^2 + (k_j - \overline{k})^2}{2}} \tag{3.3}$$

Here, A is the number of arcs of $\mathcal{N}$. $\alpha(n_i, n_j)$ is an aggregation function of the number f_i of followers of the user associated with n_i and the number f_j of the followers of the user associated with n_j. Several aggregation functions can be defined; typical examples are the maximum (i.e., $\alpha(n_i, n_j) = max(f_i, f_j)$), the minimum (i.e., $\alpha(n_i, n_j) = min(n_i, n_j)$), the sum (i.e., $\alpha(n_i, n_j) = f_i + f_j$) or the mean (i.e., $\alpha(n_i, n_j) = \frac{f_i + f_j}{2}$). In our experiments we used the latter. In this definition, as in the definition of degree assortativity expressed in Equation 3.2, k_i (resp., k_j) is the degree (meant as the sum of indegree and outdegree) of n_i (resp., n_j), while $\overline{k}$ is the average degree of the nodes of $\mathcal{N}$. Weighted degree assortativity takes into account both the degree of nodes involved in interactions and their influence, measured by the corresponding number of followers.

In this formula, the numerator represents the weighted degree covariance between pairs of connected nodes, where the weight is computed by the function $\alpha(n_i, n_j)$. Thus, the numerator is an indicator of the tendency of the influential nodes to connect with other influential nodes, taking into account both their number of connections and their number of followers. The denominator is an indicator of the variance of the degrees of nodes, weighted by their level of influence, and therefore again by their number of connections and their number of followers. It acts as a normalization factor by ensuring that the values of r_{WD} are in the real range $[-1, 1]$, where 1 denotes total assortativity, -1 indicates total disassortativity, and 0 denotes absence of both assortativity and disassortativity.

Equation 3.3 defines a more sophisticated status assortativity than that defined by Equation 3.2. In fact, for each arc of the network, it takes into account not only the degree of the corresponding nodes, but also the weight of those nodes, determined by a function that appropriately aggregates their number of followers. Thus, in this case, the differences between the degrees of the two nodes and the average degree of all nodes are weighted by the importance of the associated users. This means that a difference of the same sign (resp., different signs) of the degrees of the two nodes with respect to the average degree contributes to the assortativity (resp., disassortativity) with a weight that depends not only on the value of the difference (as was the case in Equation 3.2), but also on the weight of the corresponding users (and thus on the number of their followers).

Weighted degree assortativity is useful in cases where the degree of a node alone does not fully capture the importance of the corresponding user in the content-based network. For example, consider the case of information dissemination, where a user with few connections but many followers may be more influential than a user with many connections but few followers. In such a situation, r_{WD} provides a more holistic version of assortativity that can take into account not only users' connections, but also their degree of popularity in the network. If we consider high-influence networks, where influence plays a key role in shaping interactions, this form of assortativity proves more capable than degree assortativity in capturing and modeling the dynamics of interactions. Finally, in heterogeneous networks with a large variance in node influence, r_{WD} can help us to detect patterns that r_D might miss. This is especially relevant when we want to analyze how content diffusion or engagement patterns are related to user influence.

Having seen two forms of status assortativity, let us now define a form of value assortativity. Similar to what we said about status assortativity, the definition we will provide below is not the only form of value assortativity that can be thought of, and other forms may be defined. Since value assortativity is about user behaviors and mindsets, the topic set TS_{ij} could play a key role in its definition, and indeed the definition we propose here is strongly based on it. For this reason we call this form of value assortativity weighted topic assortativity. The corresponding weighted topic assortativity coefficient r_{WT} that we propose is as follows:

$$r_{WT} = \frac{\sum_{a_{ij} \in A} |TS_{ij}| \cdot (T_i - \overline{T})(T_j - \overline{T})}{\sum_{a_{ij} \in A} |TS_{ij}| \cdot \frac{(T_i - \overline{T})^2 + (T_j - \overline{T})^2}{2}} \tag{3.4}$$

The structure of this formula is similar to that of the formula for the computation of r_{WD}, except that the function $\alpha(n_i, n_j)$ is replaced by the cardinality of the set TS_{ij} (which represents the set of topics related to the comments from u_i to u_j, and thus refers to content), and the node degrees (which are structural measures) are replaced by the variables T_i, T_j and $\overline{T}$. All these variables also refer to content. In fact, T_i (resp., T_j) is the number of topics characterizing all messages posted by u_i (resp., u_j), while $\overline{T}$ is the average number of topics of the messages posted by all users of $\mathcal{N}$.

In this formula, the numerator is obtained by calculating the weighted mean of the products of the differences between the average number $\overline{T}$ of topics handled by the connected nodes of $\mathcal{N}$ and the number of topics handled by each pair of connected nodes in $\mathcal{N}$. The weights are given by the cardinality of the set TS_{ij} of topics in common between each pair of connected nodes in $\mathcal{N}$. This cardinality is used as a weight because it indicates how extensive the common interests are between users u_i and u_j associated with nodes n_i and n_j. Each component of the weighted mean measures the degree of similarity with which nodes engage in topics compared to the average engagement level of nodes in the network. A positive and high value of the numerator indicates that nodes with similar level of engagement in topics (both in terms of number of topics and number of common topics) tend to interact, which is an indication of value assortativity. A negative and low value of the numerator indicates that nodes with different levels of engagement on topics tend to interact, which is an indication of value disassortativity. The denominator is used to normalize the numerator. It represents the variance of the number of topics associated with each node, weighted by the number of topics in common between each pair of nodes. This configuration of numerator and denominator of r_{WT} ensures that the latter varies in the real range $[-1, 1]$, where 1 denotes total value assortativity, -1 indicates total value disassortativity, and 0 denotes the absence of any value assortativity relationship.

Equation 3.4 formalizes a version of value assortativity based on topics. The structure of this equation is similar to the one of Equation 3.3 in that, like that equation, it takes into account two factors. Given an arc connecting two nodes, the first factor takes into account the difference between the number of topics of interest to users associated with the two nodes compared to the average number of topics of interest to users in the network. It plays a similar role to the degree in Equation 3.3. The second factor is used to weight the importance of the pair of users associated with the arc and plays a similar role to the number of followers in Equation 3.3. In this case, however, the weight is given by the number of topics related to the comments posted by the first user to the second.

With respect to our definition of weighted topic assortativity, an important consideration is in order. In Equation 3.4, the set of topics relative to a comment plays a key role. We also mentioned above that a comment can contain text, images, or videos. From both a theoretical and a technical point of view, topics could be extracted from either text, images or videos. In this first paper on value assortativity, we decided to consider only topics extracted from text. This is because the techniques for extracting topics from text are more established than those for extracting topics from images and videos (as evidence, ChatGPT 3.5 did not yet implement elaboration of media content; this functionality was added in ChatGPT 4). In addition, extracting topics from images and videos requires much more computational resources and time than extracting topics from text. Finally, before we can integrate the topics extracted from images and videos into our approach, a very careful study is needed to verify that the topics extracted from text and those extracted from images and videos are homogeneous, integrable, as well as structurally and semantically compatible. All this will be the subject of our future research in this area.

3.1 Generalizability of the concepts of status and value assortativity

In the previous section, we introduced the concepts of status assortativity and value assortativity; moreover, we formalized two types of status assortativity (i.e., degree assortativity and weighted degree assortativity) and one type of value assortativity (i.e., weighted topic assortativity). We also pointed out that the concepts of status assortativity and value assortativity are extremely general and can give rise to a very large number of specifications, each of which aims to highlight a particular aspect that characterizes the status or the behavior of a user in a network. In this section, we want to explore this in more detail.

Let us start by considering the concept of status assortativity. The static characteristic considered so far is the degree centrality of nodes. However, any centrality measure (e.g., closeness, betweenness, eigenvector centralities) could be considered, and a type of status assortativity could be defined for each of them.

For example, given betweenness centrality and taking into account the weight of the users, defined in the same way as in Equation 3.3, we could define weighted betweenness centrality. The corresponding assortativity coefficient r_{WB} can be defined as:

$$r_{WB} = \frac{\sum_{a_{ij} \in A} \alpha(n_i, n_j) \cdot (b_i - \overline{b})(b_j - \overline{b})}{\sum_{a_{ij} \in A} \alpha(n_i, n_j) \cdot \frac{(b_i - \overline{b})^2 + (b_j - \overline{b})^2}{2}} \tag{3.5}$$

Here, *(i)* A is the set of arcs in $\mathcal{N}$; *(ii)* $\alpha(n_i, n_j)$ was defined above for Equation 3.3; *(iii)* b_i and b_j are the betweenness centrality values of n_i and n_j; *(iv)* $\overline{b}$ is the average betweenness centrality value among all nodes in $\mathcal{N}$. From a more intuitive point of view, recall that the betweenness centrality of a node is a measure of its ability to connect different parts of the network that are loosely connected. As in the case of Equation 3.3, Equation 3.5 considers for each arc of the network not only the betweenness centrality of the corresponding nodes, but also the weights of those nodes, determined by the followers of the users associated with them. Again, a difference of the same (resp., different) sign of the betweenness centrality of the two nodes compared to the average betweenness centrality of all the nodes in the network contributes to the assortativity (resp., disassortativity) with a weight that depends on the values of the differences and the number of followers of the corresponding users.

Status assortativity is not limited to node centralities, but can also affect other static properties of the network. For example, if we wanted to consider the clustering coefficient as the reference property, we could introduce the weighted clustering coefficient centrality. The corresponding assortativity coefficient r_{WC} can be defined as:

$$r_{WC} = \frac{\sum_{a_{ij} \in A} \alpha(n_i, n_j) \cdot (c_i - \overline{c})(c_j - \overline{c})}{\sum_{a_{ij} \in A} \alpha(n_i, n_j) \cdot \frac{(c_i - \overline{c})^2 + (c_j - \overline{c})^2}{2}} \tag{3.6}$$

Here, *(i)* A and $\alpha(n_i, n_j)$ are the same as in Equation 3.3; *(ii)* c_i and c_j are the clustering coefficient values of n_i and n_j; *(iii)* $\overline{c}$ is the average clustering coefficient value of all nodes in $\mathcal{N}$. Recall that the clustering coefficient is an indicator of the ability of a node (and thus a user) to create highly cohesive communities around itself. Assortativity based on the clustering coefficient therefore indicates whether users in a network who are able to "make community" tend to connect with other users who also tend to "make community" (assortativity), or with users who are unable to "make community" and behave as "self-centered" (disassortativity).

Value assortativity is also a general concept that can be particularized to account for different aspects of a user behaviors and mindsets. In Equation 3.4, we defined topic-based value assortativity, which takes into account whether or not a user tends to connect with other users who have similar interests. To show how multiple types of assortativity can also be defined for value assortativity, we introduce as an example sentiment assortativity, which takes into account users' sentiments as reflected in the comments they have posted. Recall that in sentiment analysis, a user's sentiment can have a value in the real range $[-1, 1]$, where -1 denotes a completely negative sentiment, 0 indicates a neutral sentiment, while 1 denotes a completely positive sentiment. In this case, however, it does not make sense to consider the weight of the nodes and thus of the corresponding users. Based on these ideas, the corresponding assortativity coefficient r_S can be defined as:

$$r_S = \frac{\sum_{a_{ij} \in A} (s_i - \overline{s})(s_j - \overline{s})}{\sum_{a_{ij} \in A} (s_i - \overline{s})^2} \tag{3.7}$$

Here, *(i)* A is the set of arcs of $\mathcal{N}$; *(ii)* s_i and s_j are the sentiments of the users represented by n_i and n_j revealed by the comments published by them; *(iii)* $\overline{s}$ is the average sentiment of the users of the network. Obviously, the value of r_S belongs to the real range $[-1, 1]$; in particular, a value of r_S equal to 1 (resp., -1) indicates a total assortativity (resp., disassortativity) of the users of the network, whereas a value of 0 for r_S indicates a lack of assortativity and disassortativity. Recall that in this case, assortativity (resp., disassortativity) indicates the tendency of a network users to connect with other users who have sentiment close to (resp., far from) her own, whereas the lack of assortativity denotes that users connect with other users regardless of the corresponding sentiments.

4 Dataset Description

As specified in the Introduction, in order to test our definitions of status assortativity and value assortativity on Threads, we had to construct an appropriate dataset, since we could not find any available open dataset that contained the Threads data that we needed for our experiments. Once we had constructed such a dataset, we decided to make it available to anyone who wanted to perform analyses on Threads. It can be found at `https://github.com/ecorradini/ThreadsDataset`. The dataset is anonymized to protect the privacy of Threads users. In this section, we describe how we obtained it and its main features.

4.1 Data Collection Methodology

To build our Threads dataset, we used a server with a 16 core CPU, 96 GB of RAM and Ubuntu 22.04 operating system. Threads allows access to its feeds in the European Union without the need for an account. To build the dataset, we followed a structured procedure consisting of the following steps:

1. *Collecting posts and comments*: we downloaded all posts and comments from the publicly available Threads feed in the European Union. In particular, we downloaded all posts and comments published in the period December 14^{th}, 2023 - February 21^{st}, 2024. Threads has a unique feature that distinguishes it from other OSNs in that each comment is, in turn, a post. For each post, we recorded some features such as its web address, the web address of the post to which it refers (if it is a comment to another post), the username of the user who submitted it, its text, the links to possible images and videos, the publication timestamp, and how many likes it received. The identity between post and comment in Threads makes it possible to record and reconstruct conversations and discussions conducted by multiple users through chains of posts/comments in which a certain post/comment represents the response for a previous one and is the reference post/comment for the next one.

2. *Collecting user data*: for each post collected, we derived some information about the user who published it. Specifically, we recorded her username, her display name, any biographical information present in her profile, any link present in her profile, and the number of her followers.

3. *Organizing data into files*: we organized all collected data into two main files, namely: *(i)* `posts.csv`, which records all data related to posts, and *(ii)* `users.csv`, which contains all data related to users. In addition, we created a special folder to store all images and videos linked by posts.

   In more detail, the file `posts.csv` has the following fields:

   - `url`, indicating the web address of the post;
   - `parent_post`, denoting the web address of the parent post, if the post is a comment; it is empty otherwise;
   - `user`, indicating the username of the user who created the post;
   - `caption`, denoting any text or caption associated with the post;
   - `image_video`, indicating the name of the visual content file in the associated folder, if the post includes images or videos; it is empty otherwise;
   - `time`, denoting the timestamp when the post was published;
   - `likes`, indicating the number of likes received by the post; it is set to 0 if the post received no likes.

   Instead, the file `users.csv` has the following fields:

   - `url`, indicating the web address of the user's profile;

- `username`, denoting the user's username;
- `display_name`, indicating the user's display name;
- `bio`, indicating the user's biographical information, if it is present in the user's profile; it is empty otherwise;
- `bio_url`, denoting web links, if they are present in the user's profile; it is empty otherwise;
- `followers`, indicating the number of followers of the user; it is set to 0 if the user has no follower.

Obtaining this dataset was not straightforward, but required a series of activities implemented through a custom scraping script. The main difficulties to overcome were related to the fact that Threads has strict security measures. The main problem we had to face in this regard was that Threads might block our IP address if we exceed the data download limits set by Threads for each IP address. To overcome this, we had to use a pool of IP addresses with an automatic system that rotated them so that each one avoided, as much as possible, exceeding the threshold of requests that could have caused it to be blocked by Threads. During each connection, we accessed Threads' public feed using Selenium and automated the scrolling process. For each post in the feed, we collected key information such as the caption, user, URL, and any associated images and videos. We also developed two specialized scrapers, one for posts and one for users. The post scraper collected more granular data, including posting time and associated comments, treating each comment as a new post. The user scraper was designed to extract all necessary information related to users. The interested reader can find the scraper code in the "`scraper`" folder of our GitHub repository `https://github.com/ecorradini/ThreadsDataset`. We note that the code may need to be updated due to constant changes in the Threads User Interface.

To ensure the quality of our Threads dataset, we conducted an assessment process, which consisted of the following tasks:

- *Removal of posts without associated users*: We examined all collected posts, identified those without users, and removed them from the dataset. In this way, we removed posts that could be orphaned or detached from the corresponding users. As a consequence, for all remaining posts, the corresponding user can be traced.
- *Removal of users who did not publish any posts*: We examined all collected users, identified those who did not publish any posts, and removed them from the dataset. In this way, all remaining users have at least one post associated with them.

4.2 Explorative Data Analysis of the dataset

Once the new Threads dataset was constructed, before using it for our assortativity experiments, we thought it appropriate to perform an Explorative Data Analysis on it so as to get an idea about its main features. The analysis was conducted on the full dataset from December 14^{th}, 2023 to February 21^{th}, 2024. Table 1 shows some basic statistics of the dataset.

An important piece of information we can deduce from this table is that most of the posts (i.e., 84%) are actually comments to previous posts. The number of posts per user is quite low and is 2.33.

Statistic	*Value*
Number of posts	129,028
Number of posts which are also comments	108,487
Number of users	55,393

Table 1: Basic statistics of the dataset

Table 2 shows some detailed statistics on the posts of the dataset. From the analysis of this table, we can see that 126,100 posts (i.e., 97.73% of all posts) have captions, i.e., they present text. Only 2,928 posts (i.e., 2.30% of all posts) have no caption. The number of posts presenting images or videos is 11,984 (i.e., 9.29% of all posts). 2,928 of these posts (i.e., 24.43%) do not have captions associated with them, while the remaining 9,056 (i.e., 75.57%) have also captions.

One thing that emerges from all these statistics is the extreme importance played by captions (and, thus, text) in the posts of Threads. In fact, only an extremely small percentage of posts do not feature text.

The appreciation level of posts within the dataset varies very much. In fact, the number of likes ranges from a minimum of 0 to a maximum of 41,064 with an average value of 56.89.

Statistic	*Value*
Number of posts with caption	126,100
Number of posts with image or video	11,984
Number of posts with caption and image or video	9,056
Number of posts without caption but with image or video	2,928
Timestamp of the first post	2023-12-14 09:01:36
Timestamp of the last post	2024-02-21 14:46:07
Minimum number of likes	0
Maximum number of likes	41,064
Average number of likes	56.89

Table 2: Some statistics on the posts of the dataset

Table 3 shows some detailed statistics on the users of the dataset. As we have seen before, each user publishes an average of 2.33 posts. These can be comments or original posts. Specifically, the average number of comments published by a user is 1.96 while the average number of original posts published by her is 0.37. Only 3,473 users, and thus only an extreme minority (i.e., 6.27% of all users), chose a username equal to their display name. Instead, the majority of users, that is 42,896 (i.e., 77.44% of all users) added a biography to their profile. 19,110 users (i.e., 34,50% of all users) added at least one link to an external site in the biographical information section of their profile. The number of followers per user varies very much and ranges from a minimum of 0 to a maximum of 9,999 with an average value of 507.38.

Table 4 presents some detailed statistics on the visual content of the posts of the dataset. As

Statistic	*Value*
Average posts per user	2.33
Average original posts per user	0.37
Average comments per user	1.96
Number of users with username equal to their display name	3,473
Number of users with a biography	42,896
Number of users with link in their biography	19,110
Minimum number of followers	0
Maximum number of followers	9,999
Average number of followers	507.38

Table 3: Some statistics on the users of the dataset

shown in this figure, there are 9,277 images and 2,706 videos in the dataset. The images in the dataset are in `.jpg` format, while the videos are in `.mp4` format. The smallest images in the dataset have a size of 320×85 pixels while the largest image in the dataset have a size of $3,072 \times 3,415$ pixels. The most common size of the images, represented by the mode, is $1,080 \times 1,350$ pixels. As for videos, the shortest of them has a duration of 0.56 seconds, while the longest one lasts 298.6 seconds. Finally, the average duration of videos is 32,97 seconds.

Statistic	*Value*
Number of images	9,277
Number of videos	2,706
Image format	`.jpg`
Video format	`.mp4`
Minimum size of images	320×85 pixels
Maximum size of images	$3,072 \times 3,415$ pixels
Most common size of images	$1,080 \times 1,350$ pixels
Minimum duration of videos	0.56 seconds
Maximum duration of videos	298.6 seconds
Average duration of videos	32.97 seconds

Table 4: Some statistics on the images and videos of the dataset

5 Investigating status and value assortativity on Threads

In the previous sections, we first introduced the concepts of status assortativity and value assortativity and then described our Threads dataset. In this section, we present the experiments that we conducted to compute status assortativity and value assortativity on Threads starting from our dataset. Once this is done, we compare our results with the assortativity values of other complex networks already computed in the past by social network researchers. Preliminary to this, however, we need to explore the network $\mathcal{N}$ associated with our dataset (see Section 3) and which is the starting point for the subsequent assortativity calculations.

5.1 Exploring the network associated with our Threads dataset

Table 5 shows some basic statistics of the network $\mathcal{N}$ associated with our Threads dataset.

Statistic	*Value*
Number of nodes	54,490
Number of arcs	89,066
Number of isolated nodes	544
Density	$2.999 \cdot 10^{-5}$
Average Clustering Coefficient	0.004041
Average indegree of nodes (excluding isolated ones)	1.65
Average outdegree of nodes (excluding isolated ones)	1.65

Table 5: Some basic statistics of the network $\mathcal{N}$

As we can see from this table, $\mathcal{N}$ consists of 54,490 nodes (users) and 89,066 arcs (connections between users). There are 544 isolated nodes; they are 1.00% of all the nodes in the network. These nodes represent users who are not connected with any other user in the network, which means that they never commented on posts by others and published a post that no one else commented on. These nodes are not useful for assortativity investigation; therefore, we decided to remove them from the network. The network density is very low and equal to $2.999 \cdot 10^{-5}$, which means that the network is very sparse. The average clustering coefficient is very low and equal to 0.004041. This result, in addition to confirming the sparseness of the network, indicates that users do not tend to form strongly cohesive groups or communities. The average indegree and outdegree of nodes are very low and are both equal to 1.65, indicating that most users are connected to 1 or 2 users in the network. The equality of the average indegree and average outdegree denotes a balance in the number of connections that users tend to make and receive.

As a next analysis, we computed the indegree and outdegree distributions of nodes in the network. It is shown in Figure 1. We computed the corresponding coefficients and obtained that $\alpha = 1.8916$ and $\delta = 0.1300$ for the indegree distribution and $\alpha = 1.3475$ and $\delta = 0.0530$ for the outdegree distribution. This result was widely expected since degree centrality is known to follow a power law in social networks [28]. This implies that in Threads there are few users who submitted many posts and received many comments and many users who submitted very few posts and received very few comments. Thus, Threads is dominated by key influencers or hubs that can significantly impact the dynamics of information flow, engagement patterns, and overall connectivity within it.

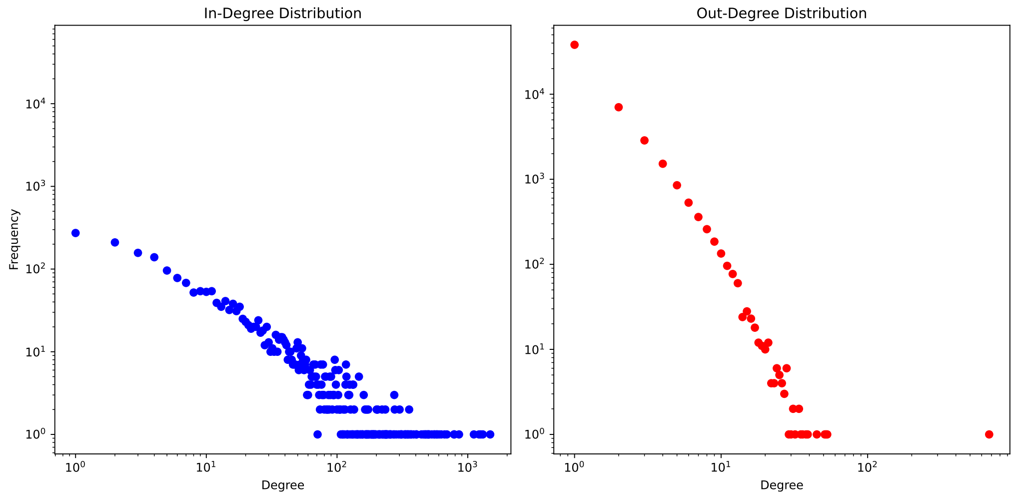

Figure 1: Indegree and outdegree distributions of nodes in the network (log-log scale)

5.2 Experiments on status assortativity

As a first analysis of status assortativity, we calculated the degree assortativity coefficient r_D (see Equation 3.2). This calculation returned that $r_D = -0.04536$. Since this value is very close to 0, we can conclude that, as far as degree assortativity is concerned, Threads users are neither assortative nor disassortative. This means that high-degree (resp., medium-degree, low-degree) users tend to connect with everyone without favoring other high-degree (resp., medium-degree, low-degree) users. This behavior differs from what generally occurs in other social networks, where researchers have generally found the existence of degree assortativity among users. This result can be explained by the fact that Threads is still a new network, and therefore there has not yet been enough time for backbones of power users, clusters of influencers, or other strong types of interactions between users with many relationships to form. Instead, at the moment it seems that the main interest of users is to distribute their content directly to as many people as possible, without thinking about building superstructures between them that would somehow facilitate an indirect transmission of content. Of course, this is not to say that in the future, as the network becomes more structured and power users and influencers emerge, Threads will not follow the dynamics of other social networks that have been around longer. Therefore, we cannot rule out the possibility that superstructures will form in the future, and that even in Threads the degree assortativity will be positive and tending to 1.

Let us now proceed with the analysis of the second form of status assortativity we have defined in this paper, namely weighted degree assortativity (see Equation 3.3). By calculating the corresponding coefficient r_{WD} on our Threads dataset, we obtained that $r_{WD} = 0.0111$. This result differs from the previous one only in sign, but again the value obtained is essentially 0. This implies that even for this second form of assortativity, Threads turns out to be neutral, i.e., there does not seem to be any significant assortativity or disassortativity relationships among network users. Recall that unlike degree assortativity, weighted degree assortativity also takes into account the weight of the nodes involved, which in our model is given by the number of followers of the corresponding users.

Consequently, this form of assortativity focuses even more on influencers. In it, the latter are evaluated both by the number of connections and by the number of followers (if we want to make a comparison, or rather a Pindaric flight, with the classical Social Network Analysis, it reminds us of the shift from degree centrality to eigenvector centrality [28]). The near-zero value for status assortativity further confirms that Threads, as a new platform, currently lacks the superstructures observed in more established social networks. This finding underscores the innovative aspect of our metrics, as they provide early insights into the evolving dynamics of this emerging platform. Another confirmation that Threads currently lacks a superstructure of users organized in a backbone or in a group to control the network or a part of it, comes from considering the value of the weighted betweenness assortativity coefficient we obtained. In fact, at the end of our experiments we obtained that $r_{WB} = 0.0428$, which once again indicates a lack of assortativity between users in Threads. In particular, it tells us that users with high (resp., low) betweenness centrality do not tend to connect only or mainly with other users with high (resp., low) weighted betweenness centrality, but, on the contrary, establish relationships with users regardless of their degree of weighted betweenness centrality.

An interesting study to better understand the relationship between degree assortativity and weighted degree assortativity is shown in Figure 2. This figure presents a scatter plot between user engagement and user influence in Threads. User engagement is determined by the number of posts a user made on Threads while user influence is determined by the number of followers the user has in this network. Examining this figure, we first observe that there is no strong correlation between the two measures. More specifically, we observe that users with low engagement values can have low, medium or high influence values. Users with medium engagement values can have medium or high influence values. Finally, users with high engagement values tend to have high influence values. On the other hand, users with low or medium influence values tend to have low or medium engagement values. Users with high influence values may have low, medium or high engagement values. These results tell us that being very active on Threads can help a user increase her number of followers, but a user could also be very influential being low active. This can be explained by considering that Threads is a new social network; therefore, a user may have "inherited" a certain influence degree, and thus a set of people following her, for her previous activity in another network. In particular, Threads has an inherent connection with Instagram, which makes it easier to inherit followers across these platforms. When a user signs up for Threads, her Threads account is linked to her Instagram account. Threads automatically adds all her Instagram followers and followings, which allows her to automatically follow users she has already followed on Instagram, and to be followed by users she has already been followed on Instagram, if they are also on Threads. This ability is very important for Threads because Instagram, with its massive user base, is a hub for influencers, content creator, and people with large followings. When these Instagram users with such characteristics join Threads, they bring part of their audience to the latter social medium. Moreover, the shared infrastructure between Instagram and Threads means that users who are recognized and followed on Instagram benefit from Instagram's reputation system, which helps them maintain their influence on both platforms. This integration suggests that an individual's influence on Threads is determined not only by her activity, or the content she posts on the platform, but also by her pre-existing activity on Instagram. Therefore, being influential with low activity can be partially attributed to existing followers from other platforms, primarily Instagram. In any case, as far as the relationship between degree assortativity and weighted degree assortativity is concerned, the fact that there is no strict correlation between engagement and influence tells us that in principle the two assortativity coefficients assess two different aspects and could, therefore, provide different results. As for Threads at this time, the values of the two assortativity coefficients are very close, but we cannot exclude that they may diverge in the future.

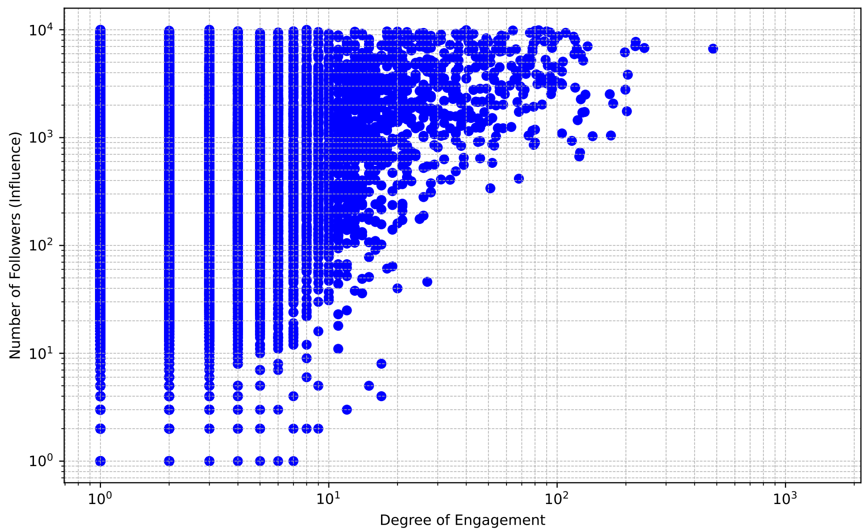

Figure 2: Relationship between user engagement and influence within Threads (log-log scale)

5.3 Experiments on value assortativity

Having evaluated the status assortativity in our Threads dataset, in this section we evaluate the value assortativity in the same dataset. To perform this task, we first needed to extract the set of topics discussed during user interactions. Given that NLP is not the main focus of this work, GPT gave us a straightforward method to obtain a topic for each post/comment. We used OpenAI's GPT-3.5 model[1]; specifically, for each post or comment in the dataset, we used `gpt-3.5-turbo` to extract the topic that best represented it. The prompt we used is:

> "Extract the main topic of discussion from the text. Just one word, not too specific, broader topic of discussion such as Technology, Health, Entertainment, Politics, etc."

More details on the API calls can be found in our GitHub repository. By doing this for all posts and comments in the dataset, we identified 846 different topics. Figure 3 shows the distribution of the 100 most frequent ones.

This figure shows that the most relevant topic is "Entertainment", which is found in posts and comments exchanged between 16,932 user pairs. The second most relevant topic is "Politics", which is present in posts and comments exchanged between 13,140 pairs of users. The figure also shows

[1] www.openai.com

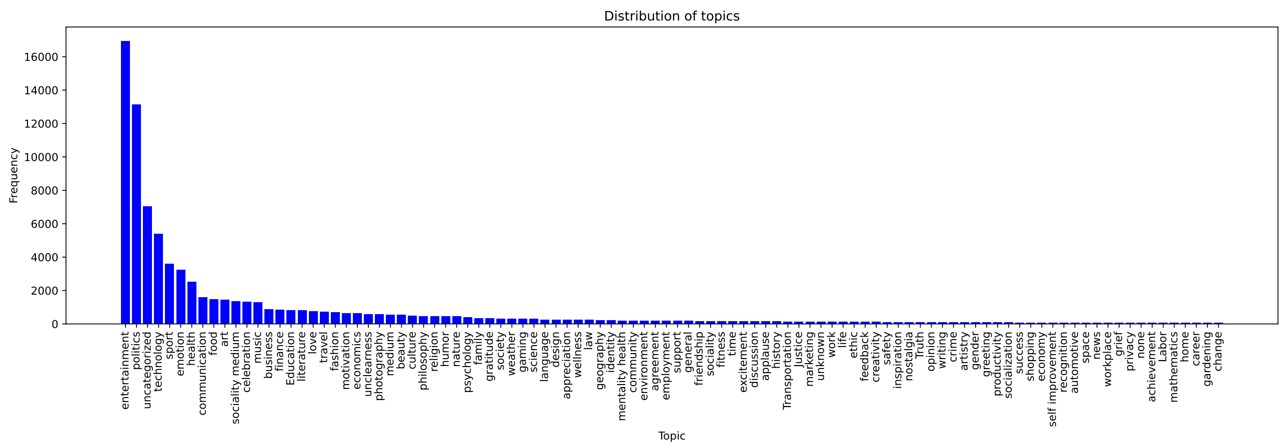

Figure 3: Distribution of the 100 most frequent topics in our Threads dataset

that there are 7,028 occurrences of an "Uncategorized" topic. This situation occurs mainly when we are in the presence of comments and posts that contain only emojis or single words such as "Yes" or "No". We did not include these comments or posts in the value assortativity calculation. Note that the distribution of topics follows a classical power law, with few topics discussed between many pairs of users and many topics discussed between few pairs of users.

With the topics and their distribution available, we were able to compute the value of the weighted topic assortativity coefficient r_{WT}, defined in Equation 3.4. At the end of the computation, we obtained a value of r_{WT} equal to 0.128. This value is positive and not zero, although it is not too high. This indicates a slight tendency of Threads users to communicate with other users who publish comments and posts on similar topics, thus showing common interests. This result suggests that there are communities or clusters in Threads whose users are interested in the same topics and tend to interact with each other. The not extremely high value of r_{WT} can be explained by considering that a user is generally interested in many topics, and therefore tends to be part of many communities by interacting with the corresponding users. Obviously, such communities only partially overlap.

An interesting final result concerns the value of the sentiment assortativity coefficient that we obtained in our Threads dataset. In fact, we obtained from our experiments that $r_S = 0.0886$. This value, similar to the one obtained for r_{WT}, indicates the existence of a slight assortativity related to the sentiment of the users, as the latter can be inferred from the comments they post. Specifically, the value obtained tells us that Threads users with a certain type of sentiment (positive, neutral, or negative) tend to relate with other users with the same sentiment as them, although this tendency is slight. One way we can explain this behavior is that users feel more attracted to users with the same "mood" as themselves, although this is mitigated by the possible existence of discussion groups where users with different sentiment come in contact, for example because users with more positive sentiment try to support those with more negative sentiment.

5.4 Evaluating the statistical significance of the results obtained

In the previous sections, we calculated the values of three status assortativity coefficients (i.e., degree assortativity coefficient r_D, weighted degree assortativity coefficient r_{WD}, and weighted betweenness assortativity coefficient r_{WB}) and two value assortativity coefficient (i.e., weighted topic assortativity coefficient r_{WT} and sentiment assortativity coefficient r_S) assumed in our Threads dataset. In this section, we want to evaluate the statistical significance of the experiments conducted to obtain these values. For this purpose, we use the Jackknife resampling technique, an approach widely used in the literature to assess statistical significance in the context of networks [23, 24]. Given a network $\mathcal{N} = \langle N, A \rangle$ (see Equation 3.1), a parameter P of interest, and the observed value OV for this parameter (i.e., the value obtained by the experiments whose statistical significance is to be assessed), the Jackknife resampling technique works as follows: It performs a number of iterations equal to $|A|$. At the i-th iteration, it removes the i-th arc from A, leaving all the other arcs unchanged, and recalculates the value of P in $\mathcal{N}$. Thus, at the end of $|A|$ iterations, there will be $|A|$ values of P available. From these values, it is possible to calculate the Jackknife Average JA and the Standard Error SE (representing, respectively, the mean and the standard deviation of the values obtained). At this point, starting from OV and SE, it is possible to construct the Confidence Interval CI for P by means of the following formula:

$$CI = OV \pm z \times SE \tag{5.1}$$

Here, CI is the Confidence Interval we want to construct, OV and SE have been previously defined, z is the critical value that determines the width of the Confidence Interval based on the confidence level we want to guarantee. Assuming that we want to guarantee a confidence level of 95% (which is the level generally adopted in experiments), z will be equal to 1.96.

We want to apply the Jackknife resampling technique to the five assortativity coefficients, which values we have previously calculated in our experiments, in order to assess whether the results obtained are statistically significant or not. We recall that the assortativity coefficient values belong to the real range $[-1, 1]$, with negative values indicating disassortativity, values close to 0 denoting neutrality, and positive values indicating assortativity. In our case, the result obtained on a coefficient can be considered statistically significant if the entire Confidence Interval obtained from its observed value by applying the formula of Equation 5.1 contains values that do not change the "polarity" from the assortativity perspective. In other words it should not be the case that some of the interval values indicate disassortativity and some indicate assortativity, and thus the interval should not contain both positive and negative values.

In Table 6 we report the observed values, Jackknife Average and Standard Error for the five assortativity coefficients of interest, while in Table 7 we report the corresponding Confidence Interval. From Table 7, we can see that the Confidence Interval values associated with the coefficient r_D are always negative; conversely, the Confidence Interval values associated with the coefficients r_{WD}, r_{WB}, r_{WT}, and r_S are all positive. We can then observe that there is no reversal of "polarity" for any coefficient, and we can therefore conclude that the experiments performed and the results obtained are all statistically significant.

Assortativity coefficient	*Observed Value*	*Jackknife Average*	*Standard Error*
Degree assortativity coefficient r_D	-0.04536	-0.04653	0.0055
Weighted degree assortativity coefficient r_{WD}	0.0111	0.0158	0.0046
Weighted betweenness assortativity coefficient r_B	0.0428	0.0467	0.0054
Weighted topic assortativity coefficient r_T	0.128	0.146	0.0067
Sentiment assortativity coefficient r_S	0.0886	0.0902	0.0045

Table 6: Observed values, Jacknife Average values and Standar Error values for the assortativity coefficients of interest

Assortativity coefficient	*Confidence Interval $CI_{95\%}$*
Degree assortativity coefficient r_D	$[-0.05614, -0.03458]$
Weighted degree assortativity coefficient r_{WD}	$[0.002084, 0.020116]$
Weighted betweenness assortativity coefficient r_B	$[0.032216, 0.053384]$
Weighted topic assortativity coefficient r_T	$[0.114868, 0.141132]$
Sentiment assortativity coefficient r_S	$[0.07978, 0.09742]$

Table 7: Confidence intervals for the assortativity coefficients of interest

5.5 Comparison with other social networks

Having computed status and value assortativity in Threads, in this section we want to compare our results with those obtained by other researchers in the past for other complex networks (especially, social platforms). In particular, we refer to the analyses presented in Section 2. Before starting our comparison, we need to recall that researchers in the past have talked about assortativity in general, without distinguishing between status assortativity and value assortativity. In what follows, we will relate the various forms of assortativity that have been studied in the past to status assortativity or value assortativity, based on the definitions of these two categories of assortativity that we have provided in this paper.

The concept of assortativity as introduced in [19] mainly considered the degree of nodes in a network and, more generally, the network structure. As a result, with respect to our taxonomy, it can be considered a status assortativity. The author has studied several forms of complex networks. The assortativity described in [22] extends the degree assortativity introduced in [19] by including direct and weighted links. All of these features concern the structure of the networks, and thus the form of assortativity introduced by the authors of this paper is a status assortativity. These authors studied several types of complex networks. The form of assortativity studied in [16] involves network configurations and thus network structure. As a consequence, in this case we are also in presence of a status assortativity. The authors considered several types of complex networks. In all these three papers, the value of the assortativity coefficients depends on the type of network; thus, some networks turned out assortative, some neutral and some disassortative.

The definition of assortativity given in [11] reproduces degree assortativity. In our taxonomy, it is a form of status assortativity. The network into examination is assortative; the corresponding coefficient ranges from 0.14 to 0.35, depending on the parameter values considered. The form of assortativity studied in [15] is based on degree correlation and is therefore a status assortativity. The authors examine several OSNs and show that some of them are assortative, some are disassortative, and some are unassortative (i.e., neutral). The authors of [31] examine nine different social networks and evaluate a form of assortativity based on social status. This is clearly a status assortativity. They show that all nine networks they examine are assortative depending on a number of parameters.

The definition of assortativity in [4] considers the happiness of users as arising from their messages. This is clearly a form of value assortativity. The authors examine it in X and conclude that X is assortative with respect to happiness, with a constant assortativity coefficient whose value is about 0.4. Similarly, the Subjective Well-Being assortativity introduced in [5] is a form of value assortativity. The authors examine it in X and conclude that X is assortative with a coefficient between 0.1 and 0.4. The definition of assortativity in [13] is based on the content exchanged and is therefore a value assortativity. The authors showed that X is strongly assortative with respect to it.

The authors of [12] analyze three different types of assortativity in Reddit. The first two types are based on degree centrality and eigenvector centrality, which are status assortativities. The authors find that Reddit is assortative for these two forms of assortativity if the corresponding centrality values belong to some intervals. In addition, they consider users' co-posting activities, examining a third form of assortativity, which is a value assortativity, and conclude that Reddit is assortative with respect to it. The form of assortativity proposed in [29] is based on the age of the users. Therefore, it is a status assortativity since the age of a user can be considered as a property of the corresponding node (similar to what we have done with the number of followers). The authors study this form of assortativity on Facebook and find that this social platform is assortative. In [7], the authors study a form of assortativity based on the membership in common online social platforms. In our taxonomy, this kind of assortativity is a status assortativity. They find that Facebook is slightly assortative, with a coefficient value of 0.074. The definition of assortativity in [9] takes into account the tendency of users of one platform to be connected to other users with whom they share the membership of other common platforms. This definition of assortativity is a form of status assortativity because it is based on the properties of nodes and their connections. The authors study it for Facebook and X and find that both platforms are strongly assortative, with an assortativity coefficient of 0.894 for Facebook and 0.675 for X.

The connectivity-based assortativity proposed in [27] is a status assortativity since it refers to the structure of the network. The authors study it on X and find that this social network is assortative, with a coefficient of 0.996. The authors of [25] study two forms of assortativity in X. The first is based on the degree of the nodes in the network, which is a status assortativity. They find that X is assortative, with a coefficient of 0.414. The second type is based on common interests and language, which is a value assortativity. Also in this case, they find that X is assortative, with a positive coefficient for all the attributes considered. The interest assortativity proposed in [8] is clearly a value assortativity. The authors apply it to X and show that X is strongly assortative with respect to three different interest categories. In particular, the corresponding assortativity coefficient is equal to 0.737 for music, 0.811 for cinema and 0.776 for sport.

6 Conclusions

In this paper, we proposed an in-depth study of the concept of assortativity introduced by Newman. First, we started from the close relationship between assortativity and the concept of homophily introduced by Lazarsfeld and Merton. In fact, assortativity can be seen in many ways as a specialization of the concept of homophily to the case of complex networks. Since there are two very different forms of homophily in the literature, namely status homophily and value homophily, in this paper we proposed to extend this taxonomy to the case of assortativity. Accordingly, we proposed a new conceptualization of assortativity and the consequent categories of status assortativity and value assortativity, providing a dual perspective that enhances the theoretical framework of assortativity. These definitions not only fills a critical gap in the literature but also offers practical tools for future research into the complex interplay of structure and content in online social dynamics.

We then applied our definitions of assortativity to Threads. This is a new social network introduced by Meta in July 2023. It is in many ways a competitor to X. To the best of our knowledge, no one has yet proposed a study of assortativity in Threads; we believe this is related to the fact that this social network was introduced very recently. Thus, a second major contribution of this paper is the evaluation of three forms of status assortativity and two forms of value assortativity in Threads. With respect to status assortativity, the result obtained is quite surprising, as it differs from similar results found for other social networks in the past, although it does not contradict them. We have also motivated the reasons for this.

In order to perform our assortativity analysis, we needed a Threads dataset storing the network users, their interactions, and the content they exchanged. Unfortunately, we could not find an existing dataset that was suitable for our purposes. Therefore, we had to build a new dataset from scratch. Then we decided to make it open and thus available to any researcher who wants to perform analyses on Threads in order to facilitate the extraction of information and knowledge about this new social platform. This is the third contribution of this paper.

The research proposed here is not only an end point, but also a starting point for further future research. In particular, the first future development could be to define a framework that allows for the definition and evaluation of new forms of status and value assortativity in social networks. This framework could also include a component that uses machine learning techniques to predict variations in assortativity based on trends that gradually emerge both in the structure of the network and in the content exchanged by its users. Next, we could extend this framework to handle other forms of dynamics within social networks besides assortativity. Another possible extension of our approach could be to examine not only text, but also images and videos in the study of assortativity. Finally, since Threads is emerging as a direct competitor to X, we would like to make a detailed comparison between these two social networks with respect to various forms of status assortativity and value assortativity. This comparison could allow us to extract knowledge patterns about the difference in both connections and user behavior in these two social networks.

References

[1] A. Al-Amery, J.L. Herrera-Diestra, Z. Du, and Z. Ertem. Epidemic Surveillance Systems and Containment Strategies in Complex Networks. *Available at SSRN 4354792*, 2023.

[2] A.F. Al Musawi, S. Roy, and P. Ghosh. Examining indicators of complex network vulnerability across diverse attack scenarios. *Scientific Reports*, 13(1):18208, 2023. Nature.

[3] M. Alfaqeeh and D.B. Skillicorn. Community detection in social networks by spectral embedding of typed graphs. *Social Network Analysis and Mining*, 14(1):12, 2023. Springer.

[4] C.A. Bliss, I.M. Kloumann, K.D. Harris, C.M. Danforth, and P.S. Dodds. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science*, 3(5):388–397, 2012. Elsevier.

[5] J. Bollen, B. Gonçalves, G. Ruan, and H. Mao. Happiness is assortative in online social networks. *Artificial life*, 17(3):237–251, 2011. MIT Press.

[6] G. Bonifazi, F. Cauteruccio, E. Corradini, E. Giannelli, M. Marchetti, D. Ursino, and L. Virgili. A Multilayer Network-Based Framework for Investigating the Evolution and Resilience of Multimodal Social Networks. *Social Network Analysis and Mining*, 14(5), 2024. Elsevier.

[7] F. Buccafurri, G. Lax, S. Nicolazzo, and A. Nocera. Comparing Twitter and Facebook user behavior: Privacy and other aspects. *Computers in Human Behavior*, 52:87–95, 2015. Elsevier.

[8] F. Buccafurri, G. Lax, S. Nicolazzo, and A. Nocera. Interest Assortativity in Twitter. In *Proc. of the International Conference on Web Information Systems and Technologies (WEBIST 2016)*, pages 239–246, Rome, Italy, 2016.

[9] F. Buccafurri, G. Lax, and A. Nocera. A new form of assortativity in online social networks. *International Journal of Human-Computer Studies*, 80:56–65, 2015. Elsevier.

[10] M. Cai, H. Luo, X. Meng, and Y. Cui. A Study on the Topic-Sentiment Evolution and Diffusion in Time Series of Public Opinion Derived from Emergencies. *Complexity*, 2021, 2021. Hindawi.

[11] M. Catanzaro, G. Caldarelli, and L. Pietronero. Assortative model for social networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 70(3):037101–037104, 2004. The American Physical Society.

[12] F. Cauteruccio, E. Corradini, G. Terracina, D. Ursino, and L. Virgili. Investigating Reddit to detect subreddit and author stereotypes and to evaluate author assortativity. *Journal of Information Science*, 48:783–810, 2022. SAGE.

[13] I. Cero and T.K. Witte. Assortativity of suicide-related posting on social media. *American Psychologist*, 75(3):365, 2020. American Psychological Association.

[14] A. Diboune, H. Slimani, H. Nacer, and K. Beghdad Bey. A comprehensive survey on community detection methods and applications in complex information networks. *Social Network Analysis and Mining*, 14(1):1–47, 2024.

[15] H.B. Hu and X.F. Wang. Disassortative mixing in online social networks. *EPL (Europhysics Letters)*, 86(1):18003, 2009. IOP Publishing.

[16] S. Johnson, J.J. Torres, J. Marro, and M.A. Munoz. Entropic origin of disassortativity in complex networks. *Physical Review Letters*, 104(10):108702, 2010. APS.

[17] P.F. Lazarsfeld and R.K. Merton. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, 18(1):18–66, 1954.

[18] R.K. Merton. *Social theory and social structure*. 1968. Simon and Schuster.

[19] M.E.J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002. APS.

[20] M.E.J. Newman. Mixing patterns in networks. *Physical Review Letters*, 67(2):026126, 2003. APS.

[21] M.E.J. Newman. Properties of highly clustered networks. *Physical Review E*, 68(2):026121, 2003. APS.

[22] R. Noldus and P. Van Mieghem. Assortativity in complex networks. *Journal of Complex Networks*, 3(4):507–542, 2015. Oxford University Press.

[23] U. Pigorsch and M. Sabek. Assortative mixing in weighted directed networks. *Physica A: Statistical Mechanics and its Applications*, 604:127850, 2022. Elsevier.

[24] S. Zhou and I. J. Cox and L. K. Hansen. Second-order assortative mixing in social networks. In *Complex Networks VIII: Proceedings of the 8th Conference on Complex Networks CompleNet 2017 8*, pages 3–15. Springer, 2017.

[25] S. Šćepanović, I. Mishkovski, B. Gonçalves, T.H. Nguyen, and P. Hui. Semantic homophily in online communication: Evidence from Twitter. *Online Social Networks and Media*, 2:1–18, 2017. Elsevier.

[26] M. Shebaro, L. N. de Moura, and J. Tešić. Improving association discovery through multiview analysis of social networks. *Social Network Analysis and Mining*, 14(1):42, 2024. Springer.

[27] V. Singh, A. Singh, D. Jain, V. Kumar, and P. Verma. Patterns affecting structural properties of social networking site "Twitter"'. *International Journal of Business Information Systems*, 27(4):411–432, 2018. Inderscience Publishers (IEL).

[28] M. Tsvetovat and A. Kouznetsov. *Social Network Analysis for Startups: Finding connections on the social web.* Sebastopol, CA, USA, 2011. O'Reilly Media, Inc.

[29] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

[30] L. Xue, P.B. Cui, and Z. Di. Identifying the structure patterns to govern the performance of localization in regulating innovation diffusion. *New Journal of Physics*, 25(5):053009, 2023. IOP Publishing.

[31] B. Zhou, X. Lu, and P. Holme. Universal evolution patterns of degree assortativity in social networks. *Social Networks*, 63:47–55, 2020. Elsevier.