Extroversion and Homophily Patterns in a Reddit Community

Leonardo Bandiera Marlia
l.bandieramarlia@studenti.unipi.it
Student ID: 606377

Alessia De Ponti
a.deponti@studenti.unipi.it
Student ID: 599725

Gian Marco Gori
g.gori21@studenti.unipi.it
Student ID: 597705

Pasquale Garofalo
p.garofalo4@studenti.unipi.it
Student ID: 681581

ABSTRACT

Previous research has documented the existence of an extroversion bias (popularity and homophily effect) in real-life social networks. In this paper, we explore these phenomena in an online social network (OSN), by studying the activity of 27K Reddit users (19K introverts and 8K extroverts) and 500K comments posted over 3 years.

We examine the presence of popularity effect using different frameworks: centrality distribution analysis, network resilience, and role discovery. The two groups present similar behaviour within the network, thus excluding this phenomenon.

Then, we evaluate the presence of homophily patterns using community detection and the conformity measure. Both approaches point to extroverts having higher heterophilic tendencies and introverts having higher homoplilic tendencies. [1]

KEYWORDS

Social Network Analysis, Attributed Networks, Role Discovery, Homophily, Community Detection

ACM Reference Format:

Leonardo Bandiera Marlia, Alessia De Ponti, Gian Marco Gori, and Pasquale Garofalo. 2024. Extroversion and Homophily Patterns in a Reddit Community. In *Social Network Analysis '24.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Popularity and homophily are structural factors of real-life social networks when considering interactions between extroverts and introverts. These characteristics cause an *extroversion bias* of the network itself: extroverts tend to have more friends (i.e. links) than introverts (popularity effect) and people tend to become friends with someone with the same extroversion level they have (homophily). This causes a general overrepresentation of extroverts in real-life social networks (Feiler and Kleinbau, 2015 [1]).

Problem Statement. We are interested in finding out whether the extroversion bias exists in online social networks (OSN) as well. Overall, we will be mainly focusing on two research questions, splitting the problem:

RQ1 Does the popularity effect exist in an OSN?
RQ2 Are there any homophilic patterns w.r.t. the extroversion level in an OSN?

Extroversion and introversion are critical behavioural aspects that are evaluated in a plethora of psychological tests. Among them, the Myers-Briggs Type Indicator (MBTI) has reached extreme popularity. MBTI is determined by a self-reported questionnaire that claims to assign to each individual their *personality type*, based on four dichotomies as reported in Table (1). The first letter of the MBTI refers to the extroversion (*E*) or introversion (*I*) of the individual.

Extroversion	**E**	**I**	Introversion
Sensation	**S**	**N**	Intuition
Thinking	**T**	**F**	Feeling
Judging	**J**	**P**	Perceiving

Table 1: Myers-Briggs dichotomies

2 DATA COLLECTION

Our objective is to build a directed, weighted network based on an online pool of users: nodes are the users themselves and one directed edge represents a comment from one user to another.

2.1 Selected Data Sources

In order to have information on the extroversion or introversion of Internet users, data was crawled from the r/MBTI subreddit on Reddit. Here users label themselves according to their MBTI, thus enabling their categorisation in *E* and *I*.

Crawling Methodology and Assumptions. Data collection has been possible thanks to https://the-eye.eu/redarcs/, an online dump with Reddit submissions and comments up to December 2022. Overall we gathered data from more than 60K users, resulting in over 1M comments spanning over 3 years (from January 2020 to December 2022).

After filtering out deleted users, users with multiple MBTI-flag or no MBTI-flag and inactive users (less than 14 days of activity), the

[1] **Project Repositories**
Data Collection: https://github.com/sna-unipi/2024_bandiera_deponti_garofalo_gori/data-collection
Analytical Tasks: https://github.com/sna-unipi/2024_bandiera_deponti_garofalo_gori/analytical-tasks
Report: https://github.com/sna-unipi/2024_bandiera_deponti_garofalo_gori/project-report

SNA '24, 2023/23, University of Pisa, Italy

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

dataset is formed by 27K users (19K for the I group and 8K for the E group) and ~ 500K comments.

3 NETWORK CHARACTERIZATION

In this section, an initial analysis will be performed on the undirected, unweighted MBTI graph, formed by 27K nodes and ~ 400K edges. The diameter of the MBTI network is equal to 7.
Table (2) compares the statistics of our MBTI graph with two benchmark models (i.e. ER and BA) having a similar number of nodes and edges. Moreover, Figure (1) shows Betweenness and Eigenvector Centrality distribution for out MBTI network and the benchmark models.

	MBTI	ER	BA
N	26809	26800	26800
L	367695	358905	375004
<k>	27.43	26.78	30.00
Components	1	1	1
ρ	0.001	0.001	0.001
C_c	0.12	0.001	0.0065

Table 2: Comparison between MBTI, ER and BA statistics

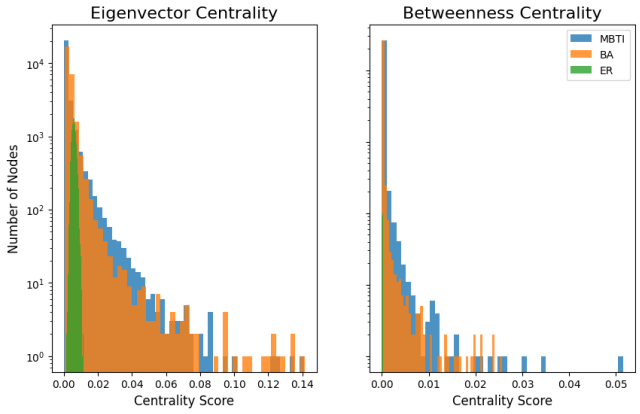

Figure 1: Betweenness and eigenvector centrality for MBTI, ER and BA

3.1 Comparison with ER

$< k >_{ER}$ and $< k >_{MBTI}$ have similar values, however, as visible in Figure (2), the MBTI degree distribution does not resemble a Poissonian, so the ER model is not well suited to represent our data.
As expected, the clustering coefficient C_c of the ER graph is much smaller than the C_c of the MBTI graph.
Moreover, we can determine the regime of the ER network:

$$< k > = 26.78 > log(N) \simeq 10.20,$$

thus the ER network is in the *fully connected regime*, in agreement with the presence of a single connected component.
The centralities distributions are very different: for the ER graph, both centrality scores have roughly the same value for all the nodes,

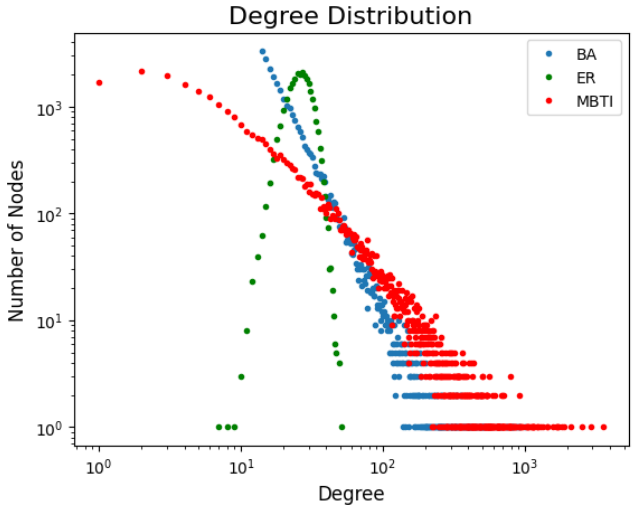

Figure 2: Degree distribution for MBTI, ER and BA

since in an random network most nodes have approximately the same degree.

3.2 Comparison with BA

Even in this case $< k >_{BA}$ is similar to $< k >_{MBTI}$, but the BA model is better suited to represent the MBTI network, since the MTBI degree distribution is positively skewed and its fat tail resembles the one of a scale free network, as can be seen in Figure (2). However, this model is not a perfect match, because when performing a best-fit using a power-law function over the MBTI degree distribution, we get $\gamma = 2.5$, whereas $\gamma_{BA} = 3$.
Furthermore, the BA C_c is also much smaller than the MBTI one. Lastly, the centrality scores are similarly distributed: they resemble a heavy-tailed power law for both the MBTI and the BA graphs.

3.3 Comparison between E and I

Considering the directed version of the MBTI network, we focus on detecting any differences between extroverts and introverts. Firstly, we look at the average in- and out-degree of each group. The results are shown in Table (3).

	$< k_{in} >$	$< k_{out} >$
E	18.79	18.29
I	16.83	17.03

Table 3: Average in- and out-degree for the two groups

Both in- and out-degree are very similar among the two groups, with them being slightly higher for the E group.
Since I nodes are twice as many as E nodes, the degree distribution has been normalised over the total number of nodes of each group. The normalised distributions are shown in Figure (3).

After the normalisation, the two groups are almost indistinguishable from the degree distribution alone.

E and I have been further compared using eigenvector and betweenness centralities as shown in Figure (4).

Neither a geometric-based score (Betweenness) nor a connectivity-based score (Eigenvector) show any significant difference among the two classes. Nevertheless, the overall highest score for both centrality measures are reached by the E label. We will further inspect this phenomenon in Section (7).

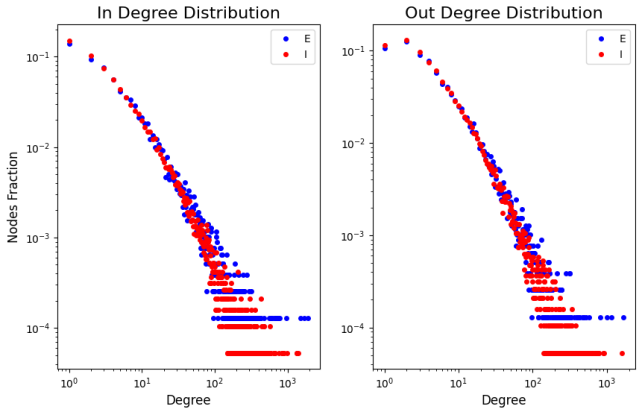

Figure 3: In- and out-degree distributions

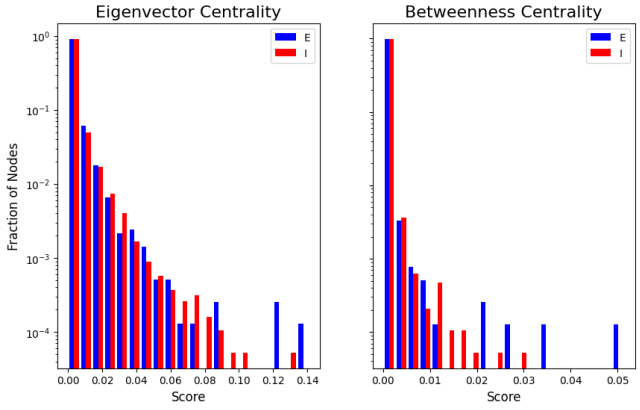

Figure 4: Eigenvector and Betweenness centrality distributions. The number of nodes is represented in log-scale

4 TASK 1: COMMUNITY DETECTION

Communities have been identified on the crawled data through a variety of algorithms, namely Louvain (LV), Label Propagation (LP), Angel/Demon (A/D), and Infomap (IM). After CD, we will evaluate the mean purity for each partition, using the external E and I labelling.

Since most CD algorithms are not implemented for directed graphs, we considered the undirected version of the graph as well. LV and LP have been applied to the undirected graph; on the other hand A/D and IM have been applied to the directed graph. The A/D

	Undirected		Directed	
	LV	**LP**	**A/D**	**IM**
# comm.	32	21	340	961
$< k >$	2.98	2.27	25.18	2.67
$< \rho >$	0.75	0.93	0.16	0.63
node coverage	1	1	0.63	1
avg. purity	0.84	0.90	0.68	0.77

Table 4: Statistics of communities with respect to various CD algorithms

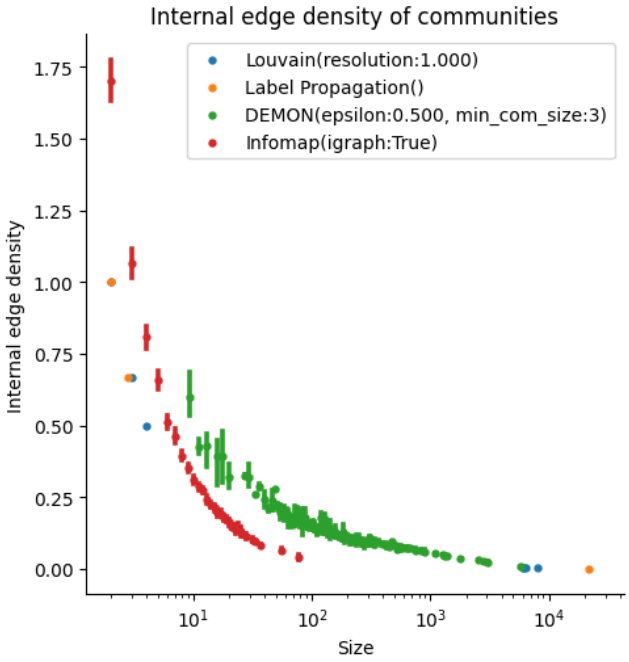

Figure 5: Internal edge density of each community of every partition with respect to the number of nodes

parameter ϵ, linked to the size of the communities that will be detected, has been set to 0.5.

A statistical analysis will compare the partitions resulting from the CD algorithms. The results are reported in Table (4). The undirected network algorithms partitioned the graph in significantly less communities than the other ones. LV, LP and IM have comparable average degree, whereas $< k >_{A/D}$ is one order of magnitude higher. The average density shows opposite behaviour: it is ~ 1 for LV, LP and IM, while $< \rho >_{A/D}$ is approximately one order of magnitude lower. The four algorithms partition the graph in mostly pure communities, with an average purity significantly above 50% in all cases. The CD algorithms for the undirected network are both equally valid, with a slight preference for LP, due to a higher internal density. IM is significantly better than A/D for the directed network.

Further research has been carried out on the mesoscale level, community-by-community.

In Figure (5), each dot represents a community being detected by the respective algorithm. In all four cases, the trend of the scatter plot is analogous: the smaller the community, the higher the internal edge density ρ. There are no small communities with low internal edge density or big communities with high internal edge density.
Among the four algorithms, the one that seems to be giving best results is IM. Indeed, LV, LP and A/D all have very large communities ($\sim 10^4$ nodes) with extremely low ρ. In contrast, Infomap produces more compact communities, ranging from a few nodes to fewer than 100 nodes, with a significantly higher internal edge density.

The violin plot in Figure (6) shows the internal edge density distribution of the four partitions. 75 % of the communities detected by LV have ρ with values between 0.5 and 1, as shown by the correspondent violin, resembling a funnel. LP and A/D present the same kind of issue: the violins are all extremely narrow but for a small area, $\rho \sim 1$ for LP and $\rho \sim 0$ for A/D. IM confirms to be the best CD algorithm for the topology of the network: the violin has variable width, being the widest with 50% communities between 0 and 0.5, but still presenting 25% with $< \rho >$ between 0.5 and 1.0.

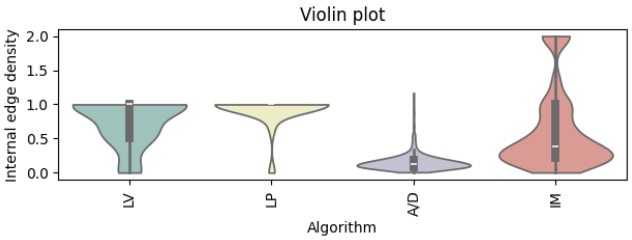

Figure 6: Violin plot: internal edge density for each algorithm

From this point we will use the partition given by the IM algorithm, not only because it applies to the directedness of our graph, but also because of the optimal parameters of the partition itself.

Using this partition, we analysed communities having a majority (> 60 %) of extroverts and introverts. Applying this threshold, 164 mixed-communities have been discarded in this part of the analysis. Results are reported in Table (5).

Communities with a majority of E are one order of magnitude less than the counterpart. This suggests a tendency to homophily for I and to heterophily for E. Moreover, E communities are on average denser than I. The average degree is comparable. I communities are significantly bigger in size on average than E communities, whereas they both have the same high mean purity.

	***E* comm.**	***I* comm.**
# comm.	75	725
$< k >$	2.33	2.74
$< \rho >$	0.99	0.56
<# comm.>	5.08	34.21
avg. purity	0.81	0.82

Table 5: Statistics of communities with majority E and I through IM algorithm

5 TASK 2: NETWORK RESILIENCE

In order to analyse the impact of strong and weak ties on the network's connectedness and resilience, it is necessary to define some measures to compute tie strength. It is possible to rank edges by *weight*, *node overlap*, and *duration of interaction* (time between the first and the last interaction). To evaluate network resilience, we implemented a progressive edge removal strategy using various ordering methods. Specifically, edges were removed in three distinct ways: random selection, descending order of edge scores, and ascending order of edge scores.
Due to computational constraints, it was essential to employ certain simplifications. We focused on a subset of the network, specifically the portion corresponding to the year 2022, which comprises 15K nodes and 150K edges. For node overlap, edge values were updated only after the removal of a finite number of edges to manage computational load effectively.

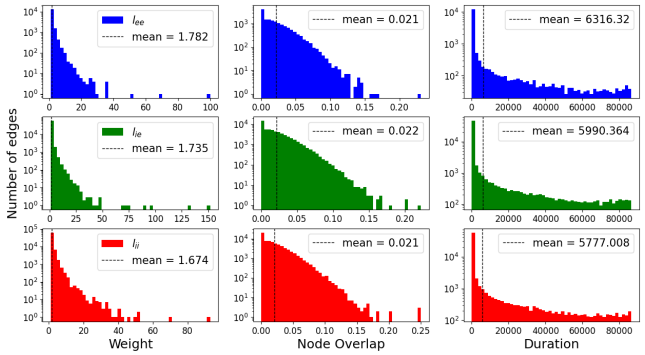

Figure 7: Distribution of edges weight, node overlap and duration

Figure (7) presents the distributions of three edge properties (Weight, Node Overlap, and Duration) for three categories of social interactions: links between extroverts (l_{ee}), links between introverts (l_{ii}), and links between introverts and extroverts (l_{ie}). Each row corresponds to a different type of interaction, while each column corresponds to a different edge property. All three types of interactions exhibit similar distribution patterns across all three edge properties.

The mean values for Weight and Duration are slightly different across the three interaction types, with l_{ee} typically having the highest means, followed by l_{ie}, and then l_{ii}. However, it is clear that differences are not large: this behaviour suggests that the nature of social interactions does not significantly change the distribution of edges properties.

Figure (8) shows the effect of different edge removal strategies on the size of the giant component of our network. The blue line, relative to random removal, represents the baseline against which other strategies can be compared. It is clear that increasing overlap removal leads to the fastest decrease in the giant component size, especially in the initial stages, suggesting that low overlap edges

(local bridges) are crucial for maintaining the network's connectivity.

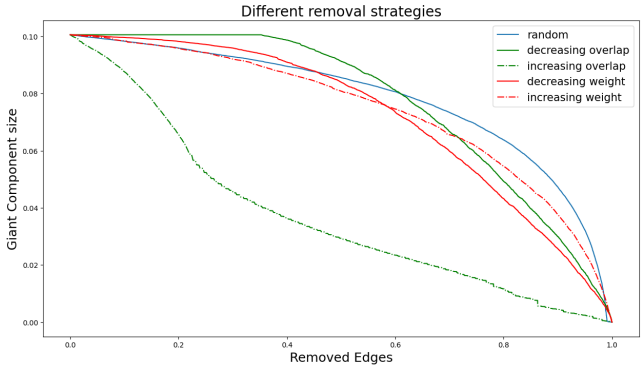

Figure 8: Effect of edge removal strategies on giant component size

In Figure (8), we omitted the curves corresponding to the duration score because they closely mirrored those of the weight score. An analysis of the correlation between the scores reveals that weight and duration are indeed correlated, as illustrated in the correlation matrix in Figure (9).

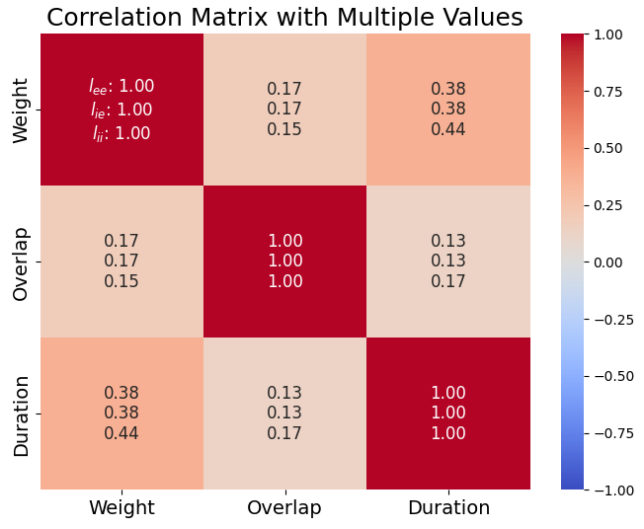

Figure 9: Correlation Matrix of edge attributes: weight, node overlap and duration

Figure (10) shows the fraction of removed l_{ee}, l_{ie} and l_{ii} in the first 30K steps, corresponding to the initial steep slope observed in Figure (8). We notice that l_{ii} interactions are eliminated at a faster rate than the other two types. This suggests that links between introverts are more likely to take the role of local bridges within our network.

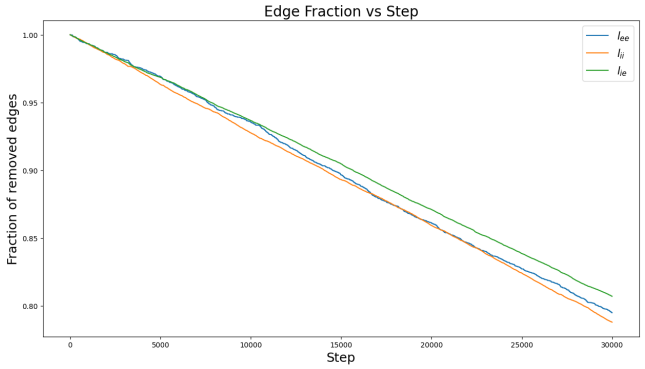

Figure 10: Fraction of removed edges in the first 30K steps of network dismantle

6 TASK 3: FEATURE-RICH ANALYSIS

The goal of this section is to evaluate homophilic and heterophilic behaviours within the network using *conformity*, a node-centric, path-aware measure of assortative mixing patterns (Rossetti and Citraro, 2021 [5]). Moreover, we will analyse possible correlations between nodes' topological attributes (namely clustering coefficient and degree centrality) and their neighbourhoods' entropy and purity.

The simplest homophily measure is Newman's *assortativity coefficient r*. For our MTBI network:

$$r = 0.073$$

which, on its own, would suggest the absence of assortative mixing patterns. In reality, the definition interval might be different from the theoretical $r \in [-1, 1]$.Moreover, Newman's coefficient might be a simplistic measure of homophily, because it flattens the whole heterogeneous context in one only score.

To further investigate potential homophilic patterns, *normalised cross-* and *intra-link scores* (η_{aa} and $\eta_{aa'}$ resp.) are evaluated. These scores are used to assess the presence of homophily or heterophily by comparing the observed number of links ($l_{aa'}$) between and within classes to those expected in a random network ($m_{aa'}$) (Park and Barabási, 2007 [3]):

$$\eta_{aa} = \frac{l_{aa}}{m_{aa}} = \frac{2l_{aa}}{n_a(n_a-1)c} \qquad \eta_{aa'} = \frac{l_{aa'}}{n_a n_{a'} c}$$

$$c = \frac{2L}{N(N-1)}$$

For our network $a \in A = \{E, I\}$:

$$\eta_{II} = 0.98 \qquad \eta_{EE} = 1.32 \qquad \eta_{IE} = 0.96$$

The η_{II} and η_{IE} scores are comparable with a random network, whereas the η_{EE} score suggests that the extrovert group might be slightly more homophilic than random.

The two approaches analysed so far (r and η) show different and seemingly confilcting results. This might be due to the fact that

these measures are very simplicistic and have different assumptions.

In order to get a more nuanced understanding of class-based connectivity, we investigate the presence of multiscale mixing patterns using a node-centric, path-aware measure: *conformity* (Rossetti and Citraro, 2021[5]).

Conformity's parameter α, also called *decay factor*, regulates the sensitivity of the score towards the distance between two nodes (i.e. $\alpha = 0 \implies$ distance plays no role, $\alpha = (>)1 \implies$ (sub-)linear decrease).

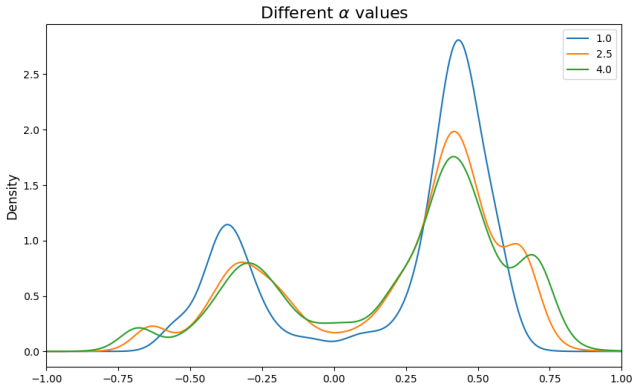

Figure 11: KDE distribution for different value of the conformity's parameter α

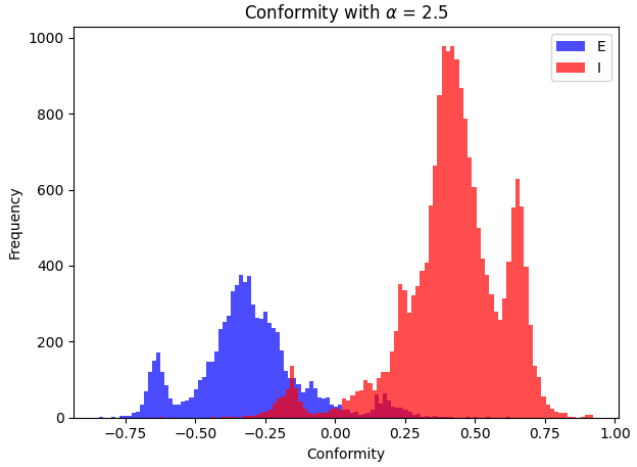

Figure 12: Conformity analysis (α=2.5)

Figure (11) shows the role of different values of α for the kernel density estimate (KDE) distribution. For all of the analysed values, the distribution is bimodal, suggesting the presence of local mixing patterns.

We further inspect the case $\alpha = 2.5$. Figure (12) suggests that introverts might have medium/high homophilic tendencies, whereas extroverts might have medium/high heterophilic tendencies. It is not possible to draw a firm conclusion on the homophilic preferences of the two groups, because we did not take into account that the population is unbalanced (I group has twice as many nodes as E group). Even though the results would require a more in-depth analysis to be fully supported, they still give an interesting insight on our problem.

Conformity suggested the presence of local mixing patterns, thus we are going to check whether some nodes' topological properties correlate with their neighbourhoods' purity and Shannon entropy.

Figure (13) shows pairplots between two nodes attributes (namely clustering coefficient and degree) and two neighbourhood measures (namely entropy and purity).

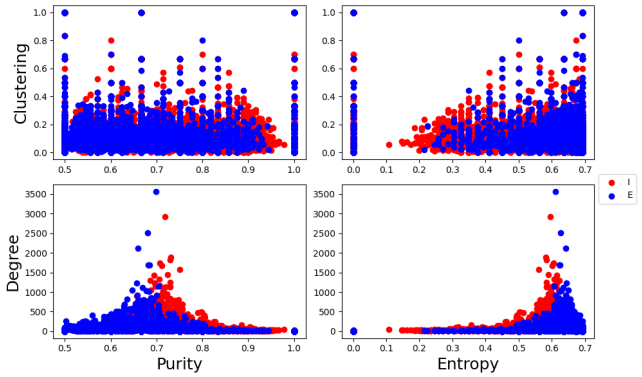

Figure 13: Pairplots between nodes' properties and their neighbourhoods' properties

.

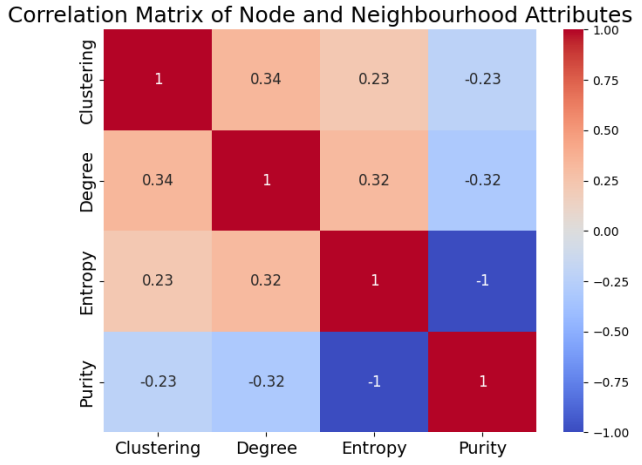

Figure 14: Correlation matrix between node attributes and its neighbourhood's entropy and purity

The Spearman correlation between these variables is shown in Figure (14). Both the clustering coefficient and the degree have non-trivial correlations with the neighbourhood's purity and entropy.

Furthermore, the correlation coefficients w.r.t purity are negative, whereas the ones w.r.t. entropy are positive. This means that higher-degree nodes and nodes in denser regions of the network tend to create links with every kind of node, regardless of their label.

7 TASK 4: ROLE DISCOVERY (OPEN QUESTION)

A *role* is a node property that groups together nodes having similar topological characteristics within the network. For instance, nodes that serve as bridges between communities, hubs with many connections, or peripheral nodes with few links can each be categorised into different roles.

We are interested in using this analytical framework to highlight possible differences in roles fulfilled by extroverts and introverts.

In this analysis we will follow closely the one perfomed in (Rafique et al., 2019 [4]). Role discovery will be performed over two scales: the whole network (*macroscale*) and network's communities (*mesoscale*), using the partition retrived by Infomap (IM) in Section (4) .

In Table (6), we set thresholds to differentiate *leaders* from other nodes for each scale.

	Macroscale		**Mesoscale**
	Degree	**Betweenness**	**Intra-Degree**
Leaders	>99.5%	>99.5%	> 99.5%

Table 6: Role Discovery thresholds

Macroscale scores are evaluated on the whole network, whereas mesoscale ones within the communities.
We identify as *leaders* nodes that have high degree, because it demonstrates the importance of a node in connecting multiple individuals. Moreover, at the macroscale level, *leaders* must have high betweenness too, as it reveals if the node acts as a bridge among other nodes or communities.

After performing the aforementioned analysis, using the thresholds in Table (6), we obtain the results in Table (7). There is no significant difference between the E and the I populations. This contrasts with real-world networks findings, in which extroversion is associated with the emergence as an informal leader (Landis, et. al., 2022 [2]).

	Macroscale		**Mesoscale**	
	E	***I***	***E***	***I***
Leaders	0.45 %	0.41 %	0.51 %	0.49 %

Table 7: Roles percentages within each group over both scales

8 DISCUSSION

In this paper we analysed the extroversion and introversion of users in an OSN through a variety of methods. Our aim is to draw parallels between real-life social networks and an OSN w.r.t. popularity effect (**RQ1**) and homophily (**RQ2**).

For (**RQ1**), we found that I are numerically overrepresented in our dataset (18K I users and 9K E users). This unbalance might be due to a preference of introverts to cultivate relationships and spend time online, as opposed to extroverts. Moreover, from a centrality distribution, tie strength and leadership role perspective, E and I are indistinguishable. This would suggest that the popularity effect does not exist in this OSN.

For (**RQ2**), we analysed assortativity with a variety of measures. Coefficients r and η did not show any significant assortativity pattern. *Conformity*, on the other hand, suggested the presence of local mixing patterns, i.e. E have higher heterophilic tendencies, whereas I have higher homoplilic tendencies. However this is not fully conclusive because we did not account for the unbalance between the two groups. Additionally, we found nontrivial correlations between nodes' attributes (clustering coefficient and degree) and their neighbourhoods' purity and entropy. At the mesoscale level, we investigated characteristics of predominantly E and predominantly I communities. E communities are denser, suggesting a higher intensity of interaction. The number of E-majority communities is one order of magnitude smaller than the number of I-majority communities. This suggests that I lean towards other I, while E are more heterophilic, thus validating the conformity observations on the macroscale.

In conclusion, our OSN does not present the same extroversion patterns as a real-life social network. On one hand, real-life social networks present popularity effect in favour of extroverts, while in our OSN there is no proof of the prevalence of one group over the other. On the other hand, while in real-life individuals tend to connect to others with similar level of extroversion, we noticed unexpected heterophilic patterns for extroverts and homophilic ones for introverts in our OSN.

REFERENCES

[1] Kleinbaum A. M. Feiler, D. C. 2015. Popularity, Similarity, and the Network extroversion Bias. *Psychological Science* 26, 5 (2015), 593–603. https://doi.org/10.1177/0956797615569580
[2] Blaine Landis, Jon Jachimowicz, Dan Wang, and Robert Krause. 2022. Revisiting Extraversion and Leadership Emergence: A Social Network Churn Perspective. *Journal of personality and social psychology* 123 (03 2022). https://doi.org/10.1037/pspp0000410
[3] Barabási AL. Park J. 2007. Distribution of node characteristics in complex networks. *Proc Natl Acad Sci USA* 104, 46 (2007), 17916–20.
[4] Wajid Rafique, Maqbool Khan, Nadeem Sarwar, and Wanchun Dou. 2019. Socio-Rank*: A community and role detection method in social networks. *Computers Electrical Engineering* 76 (2019), 122–132. https://doi.org/10.1016/j.compeleceng.2019.03.010
[5] Giulio Rossetti, Salvatore Citraro, and Letizia Milli. 2021. Conformity: A Path-Aware Homophily Measure for Node-Attributed Networks. *IEEE Intelligent Systems* 36, 1 (2021), 25–34. https://doi.org/10.1109/MIS.2021.3051291