Universität Zürich[UZH]

Blockchain & Distributed Ledger Technologies

Reproducing Structural measures of similarity and complementarity in complex networks

Alexey Buyakofu 22-737-225 Alen Frey 22-732-473 Keisuke Yokota 22-738-165 Said Haji Abukar 19-724-718

Course Name

Faculty of Business, Economics and Informatics

Date

ABSTRACT

The structure of complex networks reflects the functional properties and the formation mechanism or process. There is often a over-representation of small graphs (network motifs). The expected relationship for an abundance of triangles (3-cycles) is linked transitive relationships that arise from similarity between nodes. The paper of (**Talaga & Nowak**; **2022**) links an abundance of quadrangles (4-cycles) and dense bipartite subgraphs to the principle of complementarity. This papers follows the approach outlined in (**Talaga & Nowak**; **2022**) using their structural coefficients definition for similarity and complementarity and tries to reproduce the results and study the coefficient differences between small, medium, large offline and online social networks.

1 INTRODUCTION

1.1 Paper

The paper (**Talaga & Nowak**; **2022**) showed that an abundance of quadrangles (4-cycles) and an abundance of triangles (3-cycles) are one of the few network motifs that have an relationship to functions or properties of complex networks. The paper introduces 4-cycles as the characteristic motif for relationships that arise from complementarity or different synergies between the features of connected elements due to the similarity of one node to the neighbors of another node.

Similarity on the other hand is associated with an abundance of 3-cycles and a structural signature for transitive relationships that arise between nodes. In this context, the paper defines two so-called structural coefficients, the similarity and complementarity coefficients, as measures of the frequency of triangles and quadrangles at the level of individual nodes and edges, as well as entire unweighted, undirected graphs without loops, by generalizing the notions of local clustering and closure coefficients. Relying on the definition of quadrangles as 4-cycles without diagonal shortcuts the paper proves the association of complementarity with locally dense subgraphs of high bipartiteness, analogous to the implication of dense unipartite subgraphs by an abundance of triangles. The paper examines the coefficients and characteristic motifs for two random graph models (Erdős-Rényi and configuration model) and on multiple real biological and social networks about friendship and health advice

In an ER model, the probability of an edge is p, therefore the probability of a triangle being closed with a triple is p, and therefore the global similarity coefficient is equal to the global cluster coefficient p. A quadruple (I,j,k,l) forms a quadrangles without a cord if and only if edge (i, l) is present while edges (i, k) and (j, l) are absent; this probability is p*(1-p) and hence the complementarity coefficient is p*(1-p) and hence different from the global bipartition measures.

The configuration model is a model for studying the node coefficients and their relationships. The model enforces a certain degree sequence, the rest is as random as possible. Both coefficients are lower and upper bounded by their clustering and closure coefficients. In general, t-clustering is assumed to decrease with node degree and it has been shown that the local closure coefficient in the configuration model is positively correlated with node degree. These two results imply that structural similarity can exhibit rich, even non-monotonic correlations with node degrees depending on the structure of a

given network. As a result, the paper expected that structural complementarity should vary with respect to the degree of node in different ways that are also nonmonotonic. The theoretic expectations are consistent with the average trends observed in randomized networks fitted to the degree sequences of 28 real networks using the Undirected Binary Configuration Model (UBCM).
The results have two important practical implications.
First, the structure coefficients often tend to follow the closure coefficients for low-degree nodes and the cluster coefficients for high-degree nodes. Therefore, head triangles/quadruples dominate in low-degree nodes, and hence clustering/closure coefficients are good descriptors of triangle/quadruple density only for certain subsets of the degree spectrum. More generally, the extent of their relevance depends on the relative frequency of wedge and head paths. On the other hand, the structure coefficients are more universal because they are weighted averages of clustering and closure coefficients, where the weights reflect the relative dominance of wedge or head paths.
Second, the structure coefficients also depend on the node degrees in random graphs, so their values should be calibrated when comparing different networks based on a plausible null model such as the UBCM to to account for effects due purely to first-order structure (degree sequences).
The paper shows that there are useful domain-specific phenomena that can be used to distinguish different network types. E.g., It is shown that structural coefficients effectively distinguish between social relationships based on similarity and complementarity, with the former representing friendship and the latter representing cooperation based on advice, recognition, and skills, as indicated by an abundance of triangles or quadrangles, indicating that the latter relationships are not directly transitive

$$(i{\sim}j \wedge j{\sim}k \Rightarrow i{\sim}k) \tag{1}$$

but are second-order transitive

$$(i{\sim}j \wedge j{\sim}k \wedge k{\sim}l \Rightarrow i{\sim}l) \tag{2}$$

implying that the principle of triangle closure (2-way closure) does not capture the dynamics of such systems very well. Instead, quadrangles (3-way) closure is more appropriate. Therefore proofing that the standard assumption that similarity affects social networks and is associated with homophily and triadic closure, both of which lead to high structural equivalence between connected nodes, such that adjacent nodes are likely to have many neighbors, implying an abundance of triangles is not always justified.

Calibration and statistical significance testing of the structural coefficients were performed using the UBSCM model.

1.2 Research Question

In the paper, an open question is stated:

> Our results also point to important differences between social and biological networks. The former, with some exceptions of course, tend to be dominated by similarity while the latter are more structurally diverse, which probably reflects their heterogeneous functional properties and complex evolutionary history (we study this in more detail in "Structural diversity across the tree of life" section). However, it seems that large online social networks also feature increased complementarity relatively often (see Fig. 6A). Thus, it may be worthwhile to study differences between small and large as well as offline and online social networks in the future. In particular, to our best knowledge it is not yet clear what social processes are responsible for significantly high amounts of quadrangles in large online social networks. (**Talaga & Nowak**; **2022**)

In this project we test the stated hypothesis (marked in bold above) by comparing the structural coefficients of small and large online and offline social networks.

2 THEORY

Following for a better understanding of the following paper coefficients used and introduces in the paper (**Talaga & Nowak**; **2022**) are described.

2.1 Structural equivalence

Structural equivalence is a measure of the extent to which two nodes in a network are similarly embedded. The paper uses the definition of normalized Hamming similarity: It applies to pairs of nodes and refers to the degree of similarity of their 1-hop neighborhood. Next to structural equivalence the paper introduces two edgewise, local and global structure coefficients for similarity and complementarity.

2.2 Structural similarity

Similarity can be seen as the distance between different objects. Following the common definition of similarity in terms of distance between objects in a feature space, structural similarity is the distance between objects in a metric space, where the probability of observing a connection between them is a decreasing function of the corresponding distance. This general definition is sufficient to guarantee an abundance of triangles. A measure of the local distance of a node i to other nodes (density) at a short distance connected to it, for example, by the shortest path of length 1 (the 1-hoop neighborhood), is the local cluster coefficient of a node i s_i^w.

$$s_i^W = \frac{2T_i}{t_i^W} = \frac{\sum_{j,k} a_i j a_i k a_j k}{d_i(d_i - 1)} \tag{3}$$

T_i represents the number of triangles enclosing i, and t_i^w represents the 2-paths with i in the center (wedge triangles).
In sociological terms, it measures the extent to which my friends are friends with each other. The triadic closure includes the probability of friendship between me and my friends as well as between me and my friends' friends.
The local closure coefficient s_h solves for this.

$$s_i = \frac{4T_i}{t_i^W + t_i^H} = \frac{t_i^W s_i^W + t_i^H s_i^H}{t_i^W + t_i^H} \tag{4}$$

Here, t_i^H stands for the 2-paths (head triplets) emanating from i. Combining them, we obtain the structural similarity coefficient s_i, which is equal to the fraction of wedge and head triplets and the weighted average of s_i^W and s_i^H, implying that min(s_i^W, s_i^H) $\geq$s$_i \leq max$(s$_i^W$, s_i^H).

$$s_i = \frac{4T_i}{t_i^W + t_i^H} = \frac{t_i^W s_i^W + t_i^H s_i^H}{t_i^W + t_i^H} \tag{5}$$

s_i is a more general descriptor of local structure than s_i^W or s_i^H alone. Since s_i^W = 1 if $N_1(i)$ is fully connected, and s_i^H = 1 if there are no connections leaving $N_1(i)$, s_i = 1 if and only if i belongs to a fully connected network. Crucially, unlike local clustering and closure, structural similarity is a comprehensive measure of the density of triangles around a node i and therefore captures the full range of local structures implied by the transitivity of similarity-conditional relations. Moreover, it is defined for all nodes contained in components with at least 3 nodes. This is in contrast to local clustering, which is not defined for nodes with d_i = 1. Global similarity From a global perspective, local clustering and local clustering lead to the same conclusion, global similarity, and thus is equal to the standard global clustering coefficient and can be defined as: Where T is the total number of triangles and the denominator counts the number of triples.

2.3 Structural complementarity

Two objects are complementary if their features are different, but in a well-defined synergistic way. In contrast to similarity, for an R^k where the maximum similarity is at a minimum distance, one can always find an object that is at a greater distance and thus more complementary. Therefore, as an additional constraint, there should be only one point that is at maximum distance and thus at maximum complementarity to a node, if we consider the nodes as locations on a k-dimensional (hyper)spherical surface with $k > 1$. Therefore, there is only one node with maximal distance and for each node this distance is identical. Complementary nodes are not in each other's neighborhoods, but the 1-hoop neighborhood of i should be approximately equal to the 2-hoop neighborhood of j and vice versa, leading to an abundance of 4-cycles. Such quadruples without chores are strong quadruples.
Analogous to the similarity coefficient, the local clustering coefficient, q-clustering, measures the proportion of quadrilaterals from i compared to the 3-paths with focal nodes i (wedge quadrilaterals) compared to quadrilaterals i, and the q-clustering coefficient measures the proportion of quadrilaterals from i compared to the 3-paths emanating from I (head

quadrilaterals).

$$c_i^w = \frac{2Q_i}{q_i^W} = \frac{\sum_{j!=k} a_{ij} \sum_{k!=j} a_{ik}(1-a_{jk}) \sum_{l!=j,k} a_{kl} a_{ji}(1-a_{il})}{\sum_j a_{ij}[(d_i-1)(d_j-1)-n_{ij}]} \tag{6}$$

$$c_i = \frac{4Q_i}{q_i^W + q_i^H} = \frac{q_i^W c_i^W + q_i^H c_i^H}{q_i^W + q_i^H} \tag{7}$$

Their combination gives the structural complementarity coefficient.

$$c = \frac{4Q}{\sum_{j,j} a_{ij}[(d_i-1)(d_j-1)-n_{ij}]} \tag{8}$$

Branched networks consist of two types of distinct nodes and allow only connections between them, therefore the two types of nodes are complementary and the connection indicates high complementarity. Therefore, the complementarity coefficient is a measure of local bipartarity, but high bipartarity is not an indicator of high complementarity because density is not considered and bipartarity measures are global, not local. The global complementarity coefficient is only the fraction of quadrilaterals vs. the total number of quadrilaterals
The edge-wise similarity coefficient is the ratio of traingles enclosing nodes I and j compared to the 2-paths traversing edge (I,j), that is, the ratio of common neighbors to the total number of neighbors of I and j without themselves.

Since the similarity coefficient of node i is the average of the weighted mean of the corresponding e edge-wise similarity coefficients, and the edge-wise similarity coefficient differs from the Hamming similarity by only -2 in the denominator, it follows: Low s_i implies highly structurally equivalent neighbors, which explains the connection with transitivity. The edge-wise complementarity is the ratio of quadrilaterals containing nodes I and j compared to quadruples (j,I,k,l) and (I,j,k,l).

2.4 Social Network Analysis

Social network analysis is the study of social networks, which are networks of individuals or organizations that are connected by social relationships. There are many different measures that are commonly used in social network analysis to quantify and analyze the structure and properties of social networks. Some examples of these measures include:

2.4.1 Degree The degree of a node in a network is the number of connections it has to other nodes.

2.4.2 Centrality Centrality measures are used to identify the most important or influential nodes in a network. Examples of centrality measures include betweenness centrality, which measures the extent to which a node lies on the shortest path between other nodes, and closeness centrality, which measures the average distance from a node to all other nodes in the network.

2.4.3 Clustering coefficient The clustering coefficient of a node in a network measures the extent to which its neighbors are also connected to each other.

2.4.4 Structural equivalence Structural equivalence refers to the extent to which two nodes in a network have similar patterns of connections to other nodes.

2.4.5 Homophily Homophily is the tendency for individuals or organizations in a network to form connections with others who are similar to them in some way.

2.5 Similarity and Complementarity Measures

Both similarity and complementarity are key properties of some types of social networks, they play a crucial role in the formation of the structures present in these networks.

2.5.1 Similarity Social networks are networks of individuals or organizations that are connected by social relationships, such as friendship, kinship, or professional associations. These networks can be represented as graphs, with nodes representing the individuals or organizations and edges representing the relationships between them.

One important concept in the study of social networks is similarity, which refers to the extent to which two individuals or organizations have similar characteristics or behaviors. For example, two friends are likely to have similar interests, values, or opinions, while two organizations that belong to the same industry are likely to have similar business practices or goals.

Similarity is an important factor in the formation and maintenance of social relationships, as individuals and organizations are more likely to form connections with others who are similar to them in some way. Additionally, similarity can influence the strength and stability of social relationships, as stronger ties are often formed between individuals or organizations that are more similar to each other.

2.5.2 Complementarity In the context of social networks, complementarity refers to the extent to which two individuals or organizations have different, rather than similar, characteristics or behaviors. For example, two friends may have complementary skills or interests, such as one friend who is good at sports and the other who is good at music.

Complementarity is an important factor in the formation and maintenance of social relationships, as individuals and organizations are often drawn to others who can provide something that they lack. For example, two friends with complementary skills may enjoy working on projects together, while two organizations with complementary products or services may benefit from collaborating on a joint venture.

Additionally, complementarity can influence the strength and stability of social relationships, as individuals and organizations may feel a stronger sense of mutual dependence when they have complementary rather than similar characteristics.

3 DATASETS

We used all social datasets available from Netzschleuder, and split them into groups of online and offline social networks, and then into groups of small, medium and large networks. We only used undirected and unweighted networks. To download the datasets, we have written a script that downloads the datasets and saves them locally in the data folder.

4 METHODS

The network data was analyzed with Jupyter Notebook and Python integrated development environments. In the following section the different libraries used in this project and the paper, implmentation with the null model are discussed.

4.1 Implementation

An undirected binary configuration model (UBCM) is a (exponential) random graph model that can be used to generate networks with a given degree sequence. To fit this model to a network, you would need to first determine the degree sequence of the network, which is a list of the degrees of all the nodes in the network.
Once you have the degree sequence, you can use it to generate a random graph using the configuration model. To do this, you would first create a list of stubs for each node, where the number of stubs for a node is equal to its degree. You would then randomly connect pairs of stubs until all of the stubs have been matched, resulting in a random graph with the same degree sequence as the original network. We fitted undirected binary configuration models (UBCM) to the networks, and sampled 100 samples from the fitted models. We then calculated the similarity and complementarity measures for each sample, and compared the results to the original network. We repeated this process for each network, and calculated the p-values for each measure.

It's important to note that the configuration model is a random graph model, so the resulting graph will not necessarily be identical to the original network. However, it will have the same degree sequence, which means that it will have the same overall structure in terms of the number and types of connections between nodes. Overall, fitting a configuration model to a network can be a useful way to generate a random graph with similar structural properties to the original network.

5 RESULTS AND DISCUSSION

To ensure that here performed analysis replicates the paper results for a shared dataset, the similarity, complementarity coefficients were determined using the paper provided and to networkx adapted methods. There are minimal differences in the (A) univariate and (B) bivariate distribution of the average calibarted nodewise similarity and complementarity coefficients between this and the source paper.

In a next step the we fitted undirected binary configuration models (UBCM) to the networks, and sampled 100 samples from the fitted models. We then calculated the similarity and complementarity measures for each sample, and repeated this

process for each network. Let's first analyze if the observed increase in complementarity for large online networks is related to the size or to the online/offline character.
A look at the boxplot for the complementarity coefficient of the four types of networks indicates that complementarity does not increase with network size as it's decrease from small to large offline networks and increases from small to large online networks, the trend is also inversed for small online to small offline and large online to large offline and the in general smaller offline compared to online networks don't exhibit a lower complementarity. Therefore the paper finding is not related to network size and further analysis has to be conducted to pinpoint the cause.

Fig. 1. Global Complementarity of online and offline social networks

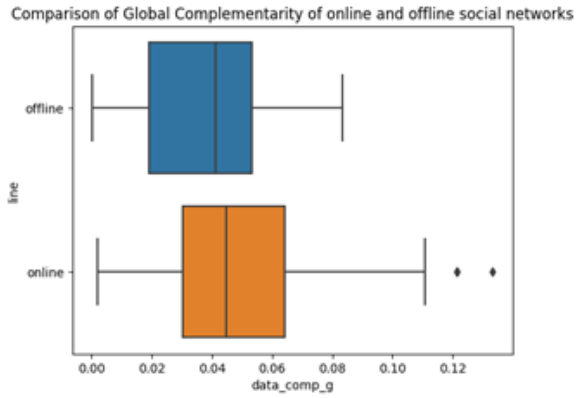


Fig. 2. Global Complementarity of small/large online/ offline social networks

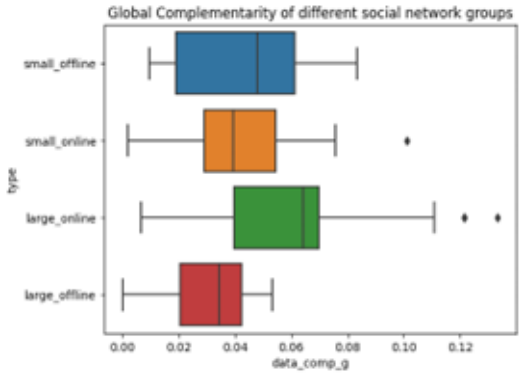


Fig. 3. Network size online social networks

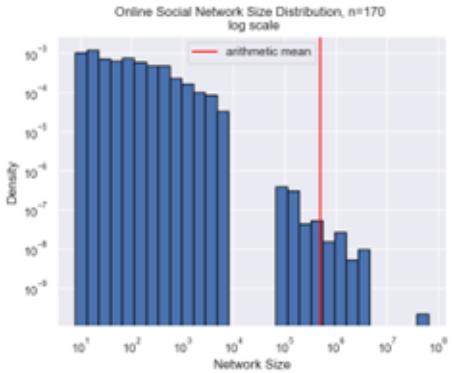
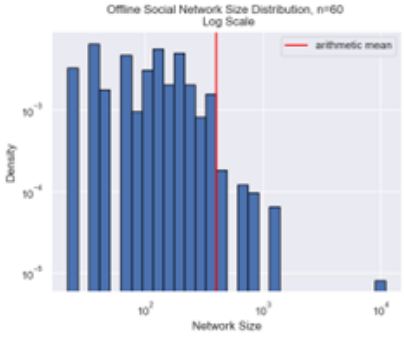


Fig. 4. Network size offline social networks



Sociologically it would be expected that online the choice of connection would lead to increased similarity in the network, this hypothesis holds true comparing the global similarity coefficient for online and offline social networks Complementarity doesn't differ significantly between online and offline networks There is no correlation between global

Fig. 5. Global similarity of online and offline social networks

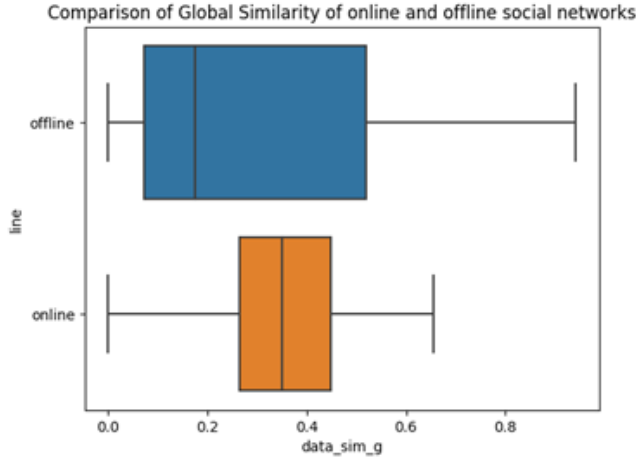
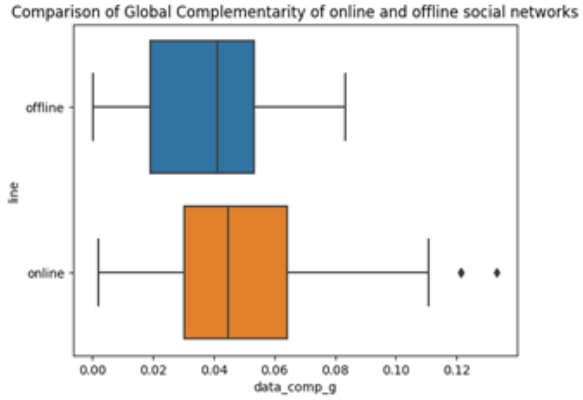
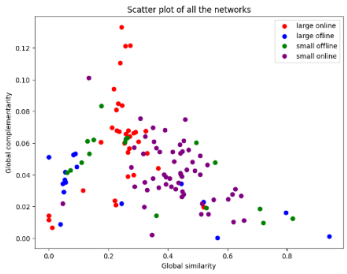


Fig. 6. Global complementarity of online and offline social networks



similarity and complementarity.

Fig. 7. Scatterplot Global similarity, complementarity coefficient of all networks



6 CONCLUSION

The aim of this project was to replicate the finding of quadrangles as network motifs for complementarity and to study the basis for the in larger online networks observed significant amount of quadrangles, indicating complementarity conversely to the expected homophily, similarity of such networks. In order to do so, we analysed four types of real world networks, small offline, small online, large online, large offline social networks by splitting up online and offline social networks available in social datasets from Netzshchleuder into small, medium large networks. The networks were fitted with the UBCM and 100 sampled, the structural similarity and complementarity coefficients were calculated. The resulted values confirm the paper hypothesis that as sociologically expected social networks sociologically justified to feature higher structural similarity then complimentarity, the paper results could be replicated. The paper assumption of increase in complementarity with online network size couldn't be pinpointed to size or online/offline character. The calibrated similarity coefficients were in general increased and the calibrated complementarity coefficient decreased compared to the null model in line with the paper findings.

7 REFERENCES

Talaga, S., Nowak, A. Structural measures of similarity and complementarity in complex networks. Sci Rep 12, 16580 (2022). https://doi.org/10.1038/s41598-022-20710-w