Complex and Social Networks
Lab Sessions Report 3
Significance of Network Metrics

Sara Montese
Victor-George Giurcoiu

UPC - FIB

19/10/2023

1 Introduction

Global syntactic dependency trees represent the hierarchical relationships between words in sentences across multiple languages, encapsulating the syntax and semantics that underlie human communication. In this study, we investigate the significance of network metrics in global syntactic dependency trees to unveil perspectives on language and network theory. Specifically, we will examine the clustering coefficient within global syntactic dependency networks. For the sake of simplicity, we concentrate on the undirected variations of global syntactic dependency networks. Here, vertices correspond to words and connections between them reflect instances of syntactic dependencies that have been observed in the context of dependency treebanks, as outlined by Ferrer-i Cancho et al [1].
The data include the description of the global syntactic dependency graphs from ten distinct languages, namely Arabic, Basque, Catalan, Chinese, Czech, English, Greek, Hungarian, Italian and Turkish. Each language adds its distinctive linguistic characteristics to enrich our comprehensive study.

2 Results

Language	N	E	$< k >$	δ
Czech	69303	257295	7.424	0.000107
Hungarian	36126	106716	5.906	0.000163
Catalan	36865	197318	10.692	0.000290
Chinese	40297	180925	8.980	0.000223
Basque	12207	25541	4.185	0.000343
Italian	14726	55954	7.599	0.000516
Greek	13283	43961	6.619	0.000498
Turkish	20409	45625	4.471	0.000219
English	29634	193078	13.031	0.000440
Arabic	21531	68743	6.385	0.000297

Table 1: Summary of the properties of the degree sequences. N is the number of vertices of the network, E is the number of edges, $< k >= 2E/N$ is the mean degree and $\delta = 2E/(N(N-1))$ is the network density of edges.

Table 1 provides an overview of the characteristics of the ten language networks. Among these networks, the Czech language network stands out as the largest, while the Basque language network is notably the smallest in size. Table 2 shows the clustering coefficients of the original networks along with the associated p-values for both the binomial and switching models. Furthermore, it provides insights about the percentage of failures within the switching model, namely the number of edge swaps that could not be performed. Table 3 presents a comparison of the four ordering of vertices (original ordering, random order-

Language	C_{WS}	p-value (*binomial*)	p-value (*switching*)	Failures
Czech	0.1217	0.0	1.0	13%
Hungarian	0.0508	0.0	1.0	11%
Catalan	0.2211	0.0	1.0	16%
Chinese	0.1708	0.0	0.0	18%
Basque	0.0467	0.0	0.1905	8%
Italian	0.1437	0.0	1.0	16 %
Greek	0.1337	0.0	1.0	15%
Turkish	0.2235	0.0	0.9048	29%
English	0.2352	0.0	0.9524	26%
Arabic	0.1885	0.0	0.3333	21%

Table 2: Clustering coefficient (C_{WS}), *p-value* using the binomial model, and *p-value* using the switching model for the ten languages. The column *Failures* refers to the percentage of random switches that could not be performed over the number of trials QE for the randomized model.

Ordering	None		Random		Ascending		Descending	
Language	p-value (binomial)	p-value (switching)	p-value (binomial)	p-value (switching)	p-value (binomial)	p-value (switching)	p-value (binomial)	p-value (switching)
Czech	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0
Hungarian	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0
Catalan	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0
Chinese	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Basque	0.0	0.1905	0.0	0.1905	0.0	0.1905	0.0	0.1905
Italian	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0
Greek	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0
Turkish	0.0	0.9048	0.0	0.9048	0.0	0.9048	0.0	0.9048
English	0.0	0.9524	0.0	0.9524	0.0	0.9524	0.0	0.9524
Arabic	0.0	0.3333	0.0	0.3333	0.0	0.3333	0.0	0.3333

Table 3: *p-values* of null models obtained by computing the clustering coefficients of only the first M vertices of the networks. We compare the results by applying different orderings of the N vertices of the graphs.

ing, increasing order by degree, and decreasing order by degree) in terms of the p-values for both the binomial and the switching models.

3 Discussion

As anticipated, it is evident that the p-values derived from the binomial model consistently fall below the 0.05 significance threshold, indicating that the clustering coefficients of the original networks are significantly different from those expected in random networks. In contrast, the p-values generated by the switching model generally exceed this threshold, with an exception observed in the case of the Chinese language network (see Table 2). A reason for the difference between the two models in terms of p-values is represented by the fact that the binomial model assumes that the network's edges are formed randomly, similar to an Erdős–Rényi random graph. Thus, it disregards the network's ex-

isting structure. On the other hand, the switching model preserves the original network's degree sequence while randomizing the edges. This means it maintains the number of connections each node has. Consequently, the reason the switching model often produces higher p-values is that it takes into account the intrinsic structure of the original network.
In terms of clustering coefficients, the results in Table 2 reveal that most languages exhibit relatively high clustering coefficients, suggesting that words in sentences within these languages tend to form clusters with more interconnected dependencies. An interesting observation emerges when we turn our attention to the Hungarian and Basque languages. Despite both having the smallest clustering coefficients among the language networks, they exhibit notable disparities in their p-values according to the switching model (the Hungarian network has a p-value of 1.0, and the Basque language has a p-value of 0.1905). Nevertheless, both p-values exceed the threshold of 0.05.
Another intriguing aspect is the percentage of failures within the switching model. The failures indicate scenarios where switching could not be performed effectively. While most languages exhibit relatively low failure rates, Turkish and English networks stand out with higher percentages of failures. These findings suggest that certain languages might exhibit unique properties in terms of syntactic dependencies that make the application of network metrics more challenging.
Regarding Table 3, it is worth noting that we obtain remarkably consistent p-values across all four orderings for both the binomial and switching models, although we are considering only M vertices. In particular, performing the optimization with the original or random ordering of the nodes takes the smallest average execution time, as shown in Table 4.

In conclusion, the significance tests under different null hypotheses provide insights into the network structures of languages. All languages exhibit clustering coefficients notably higher than the random structures according to the binomial model. However, the switching model results in clustering coefficients that are not significantly higher than random for most languages. The unique clustering patterns of Chinese suggest distinct syntactic characteristics.

4 Methods

4.1 Analysis

Let C be the true mean clustering coefficient of a given graph. Let C_{NH} be the mean clustering coefficient of a graph following a null hypothesis. We want to determine if the metric C is significantly large with regard to the null hypothesis. To do so, we estimate $p(C_{NH} \geq C)$, the so-called *p-value*, and we say that C is significantly large if:

$$p(C_{NH} \geq C) \leq \alpha \tag{1}$$

where α indicates the level of significance.
The two null hypotheses considered are:

- A binomial graph (Erdős–Rényi graph) with the same number of vertices and edges as the real network.
- A randomized graph with the same degree sequence of the original graph.

Before performing any analysis of the network properties, we remove any loop present in the input networks. Since it can be mathematically hard to find exact results, we use a Monte Carlo procedure (Algorithm 1) to estimate the p-value.

Algorithm 1: Monte Carlo Procedure for Estimating $p(C_{NH} \geq C)$

```
Data: Input parameter $C$ and the number of iterations $T$
Result: Estimated probability $p(C_{NH} \geq C)$
Initialize $f(C_{NH} \geq C) \leftarrow 0$;
for $t \leftarrow 1$ to $T$ do
    Produce a random network following the null hypothesis;
    Calculate $C_{NH}$ on that network;
    if $C_{NH} \geq C$ then
        $f(C_{NH} \geq C) \leftarrow f(C_{NH} \geq C) + 1$;
    end
end
Estimate $p(C_{NH} \geq C)$ as $p(C_{NH} \geq C) \leftarrow \frac{f(C_{NH} \geq C)}{T}$;
return $p(C_{NH} \geq C)$;
```

In order to ensure the success of this approach, we need $1/T << \alpha$. Since the procedure is computationally expensive, we set $T = 21$. With regard to the randomized null hypothesis, we use a randomization technique known as the switching model. This null model has two parameters: the original network structure and Q. The number of random switches tried is QE, where E is the number of edges. The number of trials QE has to include cases where the random switching could not be performed. We tune Q according to the coupon collector's problem, by setting $Q = log(E)$. In order to implement calculations for the switching model successfully, first we have to answer the following questions:

1. *What are the switches of edges of an undirected graph that preserve the degree sequence?*
   In an undirected graph, to perform an edge switch, we need to satisfy the following conditions:

   - edges must be non-adjacent (no nodes in common): this way we ensure that the operation maintains the degree sequence. If you switch adjacent edges, the degrees of the nodes at the endpoints of those edges would change, thus altering the degree sequence.
   - no self-loops: we want to preserve the initial structure of graphs, thus we cannot introduce self-loops

- no multiedges: we want to avoid the case of multiple edges connecting two vertices since their swap would not lead to any modification in the randomized graph.

By following these steps, you perform a switch that preserves the degree sequence of the graph. This means that the degrees of all nodes remain the same before and after the switch, and the graph structure is changed only within these constraints.

2. *What are the switches of edges of an undirected graph that preserve the degree sequence but produce edges that are not allowed?*
The switches of edges that are adjacent (that have nodes in common). If you switch adjacent edges, the degrees of the nodes at the endpoints of those edges would change, thus altering the degree sequence. For example, consider a simple graph with vertices u, v, and s, and edges $(u \sim v)$ and $(v \sim s)$. If you swap these adjacent edges, you'll end up with edges $(u \sim s)$ and $(v \sim v)$. In this case, vertex v has a self-loop, and the degree sequence has changed. Vertex v now has a degree of 2, whereas it originally had a degree of 1.

When a switching is not feasible, we proceed by trying another edge swap and continue. We do not restart the iterations.

4.2 Optimization

To reduce the computation, it is possible to estimate C_{NH} faster but with some error through a Monte Carlo procedure. A good estimate of the metric can be obtained by considering only the first M nodes, evem for $M << N$. We choose $M/N = 0.1$ We compare the results considering four different orderings of vertices:

- Original ordering.
- Random ordering of vertices (by generating a uniformly random permutation of the vertices).
- Increasing order by degree.
- Decreasing order by degree.

The results after applying the optimization technique above are shown in Table 3. In Table 4 we compare the average execution time of computing the mean local clustering coefficient for different node orderings. The metrics refer to the optimized binomial model. It can be seen that ordering nodes by *increasing* and *decreasing* degree leads to the same results of the other orderings, but is generally at least 10 times slower.

Furthermore, when exact calculations or good approximations are computationally intensive, the p-value can be bounded in an analytical way. An idea could be to establish a bound based on Chebyshev's inequality. *Chebyshev's*

Ordering	None	Random	Increasing	Decreasing
Language				
Czech	4.0	4.7	146.3	99.5
Hungarian	2.2	2.5	46.8	30.1
Catalan	3.2	4.6	87.8	62.9
Chinese	3.3	3.8	86.5	59.9
Basque	0.4	1.0	4.5	2.8
Italian	0.8	1.2	9.7	6.6
Greek	0.7	0.9	9.4	6.8
Turkish	0.9	1.1	11.7	7.2
English	3.2	4.0	64.1	47.6
Arabic	1.2	1.4	17.6	11.7

Table 4: Average execution time (in seconds) of computing the mean local clustering coefficient for different node orderings. The metrics refer to the optimized binomial model.

inequality for the range variant guarantees that within a specified range or distance from the mean, no more than a specific fraction of values will be present. For any random variable X with mean μ and standard deviation σ, and for any positive constant a, it states that:

$$P(\mu - a\sigma \leq X \leq \mu + a\sigma) \geq 1 - 1/a^2 \tag{2}$$

We can think of:

- μ as the mean local C_{WS} of the original sample.
- σ as the standard deviation of the original sample.
- X as the local C_{WS} of the samples from the null hypothesis.
- 1 - $1/a^2$ as the lower bound on the probability.

Nonetheless, it is important to notice that these bounds are conservative, meaning they provide a worst-case scenario for the p-value. Additional research is required to confirm whether this approach is correct and advantageous.

4.3 Implementation

Our decision to use Python for the implementation of this analysis is primarily based on our lack of expertise in C or C++. Another reason for the choice of this programming language is given by the presence of the package *networkx* [2], a tool we are already familiar with.

References

[1] Ferrer-i Cancho et al. “Patterns in syntactic dependency networks”. In: *Physical Review E* (2004).

[2] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. “Exploring Network Structure, Dynamics, and Function using NetworkX”. In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, 2008, pp. 11–15.