Individual node's contribution to the mesoscale of complex networks: A modified implementation

Wilmer Uruchi

January 19, 2020

Abstract

In the following work we explore the research done in [4] about measuring the contribution of individual nodes to the mesoscale of complex networks. In the original work they use undirected networks, we apply the necessary changes to adapt the approach to directed networks and measure the effect of the out and in degree separately.

1 Introduction

Representing a real system as a network allows for the abstraction of some key concepts and the study of the components of the system using the properties of network models, and the mathematical tools that they provide. In this work we take the model proposed in [4] that was originally aimed at understanding the biological function of observed anatomical structures, and we apply it to analyze twitter users interaction for a given period of time and topic.

We propose some little changes to the original model because our subjects of study have a directed nature and, conversely, those studied in the original paper are more on the undirected spectrum. Consequently, some changes are implemented in the mathematical formulation of the metrics studied in the model. Furthermore, a new set of experiments is added to try these changes.

The idea for this work started as an study by Universitat of Barcelona physicists who tried to identify some key roles in the communication structure of the users of twitter during specific social events, i.e. strikes, political unrest, or even just a plain day. That work was based in metrics that have been improved in subsequent research. This work is an attempt at analyzing similar situations than those in the original study with the improved formulation of the metrics, and also add the new dimension of dealing with directed networks.

2 Characterization of a node's role in a modular network

In this chapter we go very quickly through the mathematical basis of the model. Further details can be found in the original paper [4]. We will put special emphasis into the changes we applied for our implementation.

2.1 Community Finding

For the purposes of community finding, we make use of Infomap [1], a network clustering algorithm based on the map equation [2]. Infomap is well suited for directed networks, although its options allow for experimentation with a wide array of network types.

Community finding will always be performed with the original directions of the edges in the dataset.

2.2 Hubness Index

Hubs are typically defined as those nodes with many more connection than others, in this section we formalize the calculation of the index that determines this characteristic.

The main idea of this calculation is to compare the degree of a node in a community, to the degree this node would have in a random generated graph with the same density as the sample.

We define hubness of a node in a network of size N and density ρ as:

$$h_i = \frac{k_i - \langle k \rangle_R}{\sigma_R} = \frac{k_i - (N-1)\rho}{\sqrt{(N-1)\rho(1-\rho)}}$$

Where $\langle k \rangle_R$ is the mean out-degree of the equivalent random graph against which the out-degree k_i is compared. The hubness is negative $k_i < \langle k \rangle_R$ for nodes less connected than expected from randomness. In general, the greater the value of the hubness index, the more confidently we can state that the referenced node is a hub.

We characterize nodes by two hubness values: local hubness h^l and global hubness h^g, where h^g is the hubness in relation to the network and h^l is related to the community to which the node belongs. For h^l, $N^{'}$ and $\rho^{'}$ take their values from the community.

On further experimentation, and expanding the definition, we use the in-degree for k_i.

2.3 Participation and dispersion indices

The **participation vector** P_i is a vector whose elements P_{im} represent the probability that node i belongs to community C_m, where $m = 1, 2, ..., M$. The probability if defined by: $P_{im} = \frac{k_{im}}{N_m}$, where N_m is the size of the community, k_{im} **assumes the value of the out-degree and is changed to the in-degree in further experimentation**. Participation vectors are normalized such that $\sum_{m=1}^{M} P_{im} = 1$. As an example, if $M = 4$,

and node i belongs only to community 1, then $P_i = (1, 0, 0, 0)$. If node were connected to all communities, then $P_i = (1/4, 1/4, 1/4, 1/4)$.

Then we define the **participation index** $p_i = 0$ if the node devotes all its nodes to a single community, and $p_i = 1$ if it is equally connected among all communities.

$$p_i = 1 - \frac{\sigma(P_i)}{\sigma_{max}(M)} = 1 - \frac{M}{\sqrt{M-1}}\sigma(P_i)$$

The larger the p_i the more difficult is to classify it into a single community.

Then we define the dispersion index d_i in a similar way than p_i but only considering non-zero entries.

$$d_i = 1 - \frac{\sigma(P_i^{'})}{\sigma_{max}(M^{'})} = 1 - \frac{M^{'}}{\sqrt{M^{'}-1}}\sigma(P_i^{'})$$

The **dispersion index** is a measure of how difficult is to classify a node. Furthermore, not all combination of dispersion and participation are possible:

$$p^{+}(d, M) = 1 - \sqrt{\frac{M}{M-1}\left[\left(\frac{2-d}{2}\right)^2 + \left(1 - \frac{2-d}{2}\right)^2 - \frac{1}{M}\right]}$$

2.4 Application to twitter datasets

2.4.1 Implementation

The formulas described in the paper have been implemented using C++ in the following way:

1. The program reads an ".edge" file and transforms the input into a Graph object defined for this implementation.

2. The Graph edge is processed and a link-list formatted file is created.

3. The link-list file is used as input for Infomap.

4. Infomap is called as a command (This implies that an installation of Infomap already exists), the link-list file is sent as input, as options we define "-N 5 –directed" so the algorithm runs 5 iterations and assumes the input describes a directed network.

5. Infomap runs independently and produces a ".tree" output with the details of the results of the community partition.

6. The program reads the ".tree" file and starts to calculate the individual metrics mentioned in the previous sections, starting with the **participation vector** and ending the global and local hubness.

7. The result is 2 ".graphml" that store all the calculated metrics as node attributes.

8. Finally, analysis is performed using R tools.

All the code and results can be found in [3]. It will be continually updated with further implementation as the research project continues.

2.5 Data

The data consists of three datasets:

- Marta_Rovira: A dataset that shows some community structure.
- Vaga8Nov: A dataset that shows strong community structure. There are three communities that include the majority of nodes.
- nochebuena: A dataset with weak community structure. Serves a null model.

2.5.1 Results

In this section, when we refer to some graphic or result as out-degree it is because the corresponding metric has been calculated using the out-degree of the node; consequently, in-degree refers to results using the in-degree of the node.

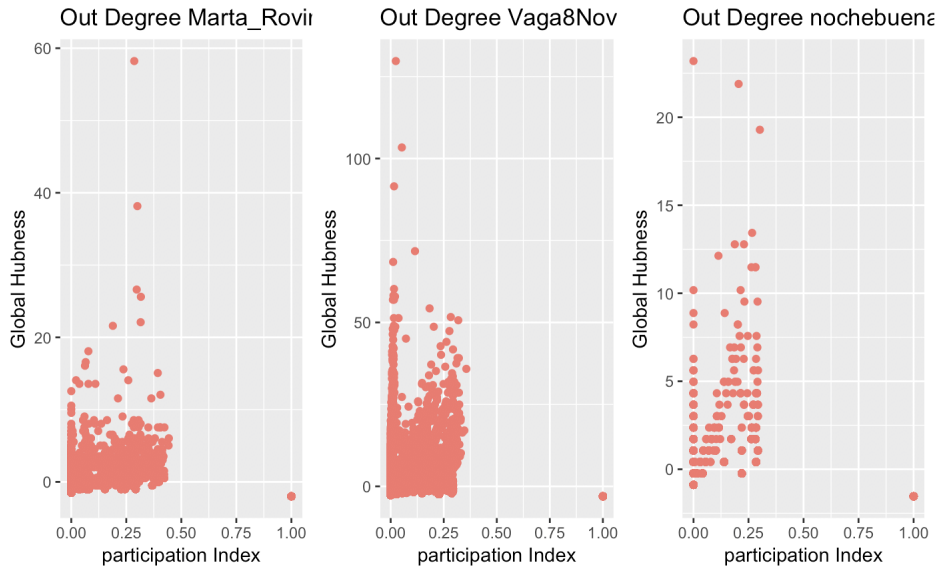

Figure 1: Participation Index compared to Global Hubness for Out-Degree

These are two of the most useful metrics calculated. Global Hubness h^g tells which nodes have a higher connection in the whole network, and is compared to Participation Index, which tells us how distributed is the degree of a node among communities, $p_i = 0$ if the edges of a node are located in a single community.

We observe that "Marta_Rovira" presents a high concentration of nodes with low h^g and low p_i. These metrics give some intuition into what happened during the timespan this data was retrieved, i.e. participants tended to concentrate their interactions with members of their communities. The points h^g represent those participants that referenced through retweets or comments to a great quantity of other participants.

"Vaga8Nov" shows a similar distribution than "Marta_Rovira", but in this case p_i is even lower, suggesting a higher quantity of participants or nodes that did not choose to interact with other nodes that did not share their opinions.

"nochebuena" shows a similar trend, however we know that this dataset does not present strong community structure, so we can attribute the low p_i to the fact that there are not big enough communities that would make this index grow in a way that will make it comparable to the results from the other datasets.

The three plots show a cluster of points with $p_i = 1$. This is due to the fact that the definition of the calculation of this metric does not include the case where the out-degree is 0, same for the in-degree. So, to avoid a division by 0 scenario, that code just returned 0 in those cases. We later concluded that it was best to give NA as a result for the p_i in this cases.

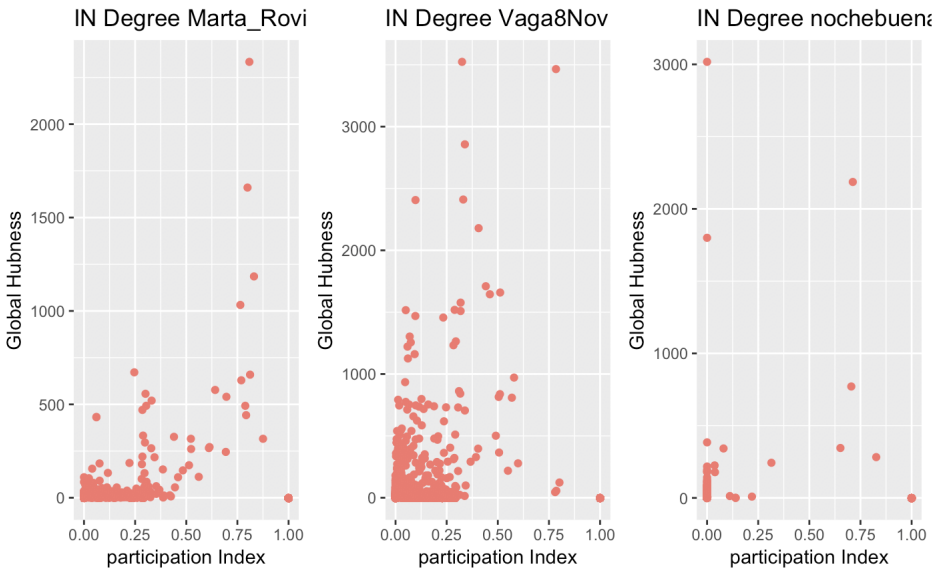

Figure 2: Participation Index compared to Global Hubness for In-Degree

In this case we are looking at the in-degree results of p_i vs h^g.

We see in "Marta_Rovira" that the scale corresponding to h^g increases significantly, suggesting that there are nodes with high in-degree. Moreover, p_i signals that the in-degree of the nodes seems to come from a more varied than the out-degree. We can say that people tend to reference more people outside their community, and as a consequence, they receive less references from people in their own community.

"Vaga8Nov" seems to follow the trend described by "Marta_Rovira".

On the other hand, in "nochebuena" shows the results corresponding to a network with weak community structure, a low p_i and low h^g. The node with high h^g corresponds to an event.

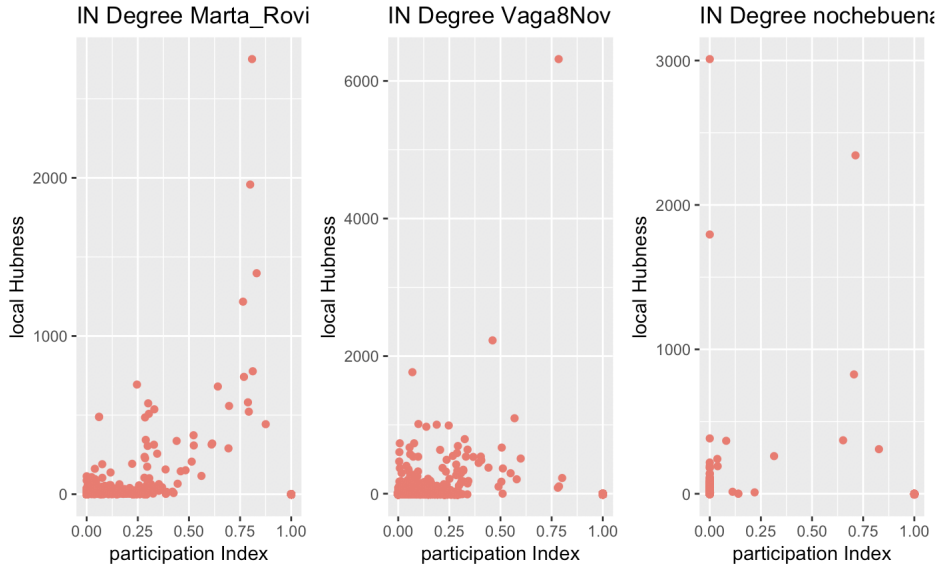

Figure 3: Participation Index compared to Local Hubness for In-Degree

In "Marta_Rovira" the results looks almost identical to that of h^g; however, the change in the scale suggest that the node that showed higher h^g in Figure 2 has an even higher h^l suggesting that this node has high in-degree outside his community and comparable higher inside his community. The results in "Vaga8Nov" and "nochebuena" suggest a similar trend.

2.6 Comments

Although the paper [4] presented a definition of metrics aimed at undirected networks, there is a reference to an experiment performed comparing the results using the in-degree and out-degree of the same network. We used the description of that experiment for our

own experimentation. However, it is important to mention that the community finding is always performed using the out-degree.

The fact that the definition of p_i does not include cases where the out or in-degree is 0 is a consequence of these definitions being developed for undirected graphs. We concluded that the best workaround that $p_i = NA$ when the degree is 0, since it seems to be most logical sound assigned value.

The work and experiments performed as part of this project are merely the first step in a much wider and ambitious project that aims at using the metrics presented in this report to divide nodes into roles that best describe the nature of the participation of people in interactions done through twitter during relevant social events. This might result in some changes in the definitions of these metrics or even the introduction of new ones.

The implementation was done using C++, this language allows for efficient performance and since most of the calculation of metrics is independent among nodes, the opportunities for parallelization are plenty. The larger experiment consisted of approximately 100k nodes and 450k edges. Processing this amount of information took around 20 min, that included reading the input, processing Infomap, processing Infomap result, metric calculation, and output writing.

The repository [3] also includes the file "/code/analysis/data_process.Rmd" with some R code to process the output (.graphml) generated by the program. The output format also allows for other network specialized software to easily read the output and process it.

References

[1] `https://mapequation.github.io/infomap/`

[2] M. Rosvall, D. Axelsson, C.T. Bergstrom *The Map Equation* Eur. Phys. J. Special Topics 178 13 - 23, 2009.

[3] `https://github.com/wuruchi/commub`

[4] Florian Klimm, Javier Borge-Holthoefer, Niels Wessel, Jürgen Kurths, Gorka Zamora-López *Individual node's contribution to the mesoscale of complex networks*, New Journal of Physics 16, 2014.