Centrality is All You Need

Cristiano Colpo[†], Matteo Piva[†]

Abstract—**In the last period mass media and social networks have acquired more and more relevance in the political field. In this report we deal with Facebook activities of politicians collected in November 2017, where nodes are politicians and edges are the mutually likes among the pages. We exploited an analysis of the features of the network at node level and graph level in order to retrieve useful information about the network. The main result we found consist in the fact that there are some politicians that are apparently more central in the network, given the fact that they have high values for the majority of the features computed. Finally an analysis of the significance of the results obtained is performed considering random graphs.**

Index Terms—**Politics, social media, social network, facebook, graph analysis, centrality measures, clustering coefficient.**

I. Introduction

Within the last thirty years, political systems experienced a continuous evolution, at least for all the democratic countries [1]. Mass media become more and more present in political domain, assuming sometimes a character of necessity. However, even if they could represent a positive change for democracies around the world, they are not just passive medium for communicate political content, but they have their own aims, and rules and they must be used carefully.

Nowadays, we are moving from traditional media to a new kind of media: social media. Thanks to the diffusion of social network sites like Facebook (2004) or Twitter (2006) [2], social media have been established as one of the most popular Internet services in the world. For example, for the U.S. presidential election of 2008, the presidential candidate Obama created a blog to communicate better with new generations, and recruit campaign volunteers across all the country. This new way of transmitting information shows to meets the requisite of the public sphere [3]. In fact, the Internet allows unlimited access to information and equal participation [4]. From posts to comments and likes, from tweets to retweets everyday every citizen with an Internet connection could publish its own opinion and perspective on the online or keeping a discussion with everyone else around the world.
Facebook in particular is the most used social network, with 59% of the world's Internet users, 2.80 billions of active users monthly and 1.84 billions every day. Here citizens could follow politician pages through a "Like", monitor the political agenda and interact with them in the political discussion. On the other side, politicians from their pages could expose their political ideas and plans, organize public event, live directs, and understand the support of the electorate.

In this context of interlaced relationship, network science, and in particular the social network analysis branch are used to study and evaluated the connections. In this field citizen, politicians, groups or organization are the nodes and the interaction among them the link of a complex network which is a network of related entities. This science allows to map and measure all the present relationships to understand the information flow, the interacting parts, who knows whom, and who shares what information with whom [5]. In this way it is possible to identify the most influent individuals or groups which play a central roles in the politic scenario, and find out how the different nodes are organized inside the network.

II. Research question

The central research question of our project is highlighted in the project title: "Centrality". In fact we want to find which are the most influential political figures, based on the mutually liked Facebook pages.
In general, an influential politician should be represented by a node with high degree, because lots of

[†]Department of Information Engineering, University of Padova, email: cristiano.colpo@studenti.unipd.it (2017898), matteo.piva.6@studenti.unipd.it (2020352)

other nodes (more or less important) are connected with it, high closeness centrality, due to the fact that the more dominant is a node, the closer he is to all other, and high betweeness centrality because it captures how much a node is in-between others. Thus, we want to use degree, closeness centrality and betweeness centrality (the last two both and approximated), to detect the most dominant politicians and to observe how the ranking change using the three different metrics. Then, we use local and global clustering coefficients to asses how nodes (and so politicians) tends to be organized among them in the network.
Finally, we want to perform a significance check of the features computed by comparing them with random graphs. This is done through z-score and p-value using implementations of Erdös-Renyi-Gilbert and Chung-Lu random graphs models.

III. The Database

The database we use is fb-pages-politician [6], [7]. Data are collected from Facebook in November 2017, and nodes represent the politicians pages (blue verified Facebook page networks of different categories are considered), while edges are the mutually likes among the pages. This it an undirected graph and it is connected. The main characteristics of the network are reported in table 1, while the network itself is shown in Fig. 1.

Parameters	Value
Nodes	5.9K
Edges	41.7K
Density	0.00239013
Maximum degree	323
Minimum degree	1
Average degree	14
Number of triangle	523.9K
Average number of triangle	88
Maximum number of triangle	3.7K
Average clustering coefficient	0.385096

TABLE 1: fb-pages-politician network parameters.

IV. Methods

To compute our features of interest, we used the implementations of methods available on the module Networkit. Both exact and approximated versions were used where possible. We also did some tests with NetworkX in order to compare the results in some cases, but in general the time required by the methods of NetworkX is higher, so we decided to use mostly Networkit. The following are the features/methods used:

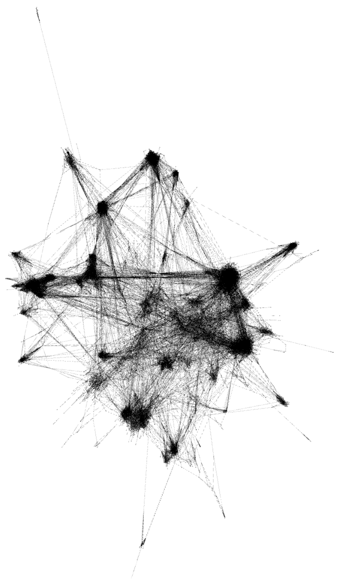


Fig. 1: Politicians network

- Diameter: Represents the maximum distance between two nodes on the network.
- Effective Diameter: Calculates the effective diameter of a graph. The effective diameter is defined as the number of edges on average to reach a given ratio of all other nodes. We decided to try both exact and approximated methods.
- Degree: It the number of nodes to which a node is connected. It is useful to have an overview of the degree distribution on the network.
- Closeness Centrality: Constructs the Closeness class for the given Graph G. If the Closeness scores should not be normalized, set normalized to False. The run() method takes $O(nm)$ time, where n is the number of nodes and m is the number of edges of the graph.
- Approximated Closeness Centrality: Approximation of closeness centrality according to algorithm described in Cohen et al., Computing Classic Closeness Centrality, at Scale [8]. The algorithm approximates the closeness of all nodes in both directed and undirected

graphs using a hybrid estimator. First, it takes nSamples samples (after some experiments we selected 100 as number of samples). For these sampled nodes, the closeness is computed exactly. The pivot of each of the remaining nodes is the closest sampled node to it. If a node lies very close to its pivot, a sampling approach is used. Otherwise, a pivoting approach is used. The epsilon parameter (we choose the value 0.05) is used for the error guarantee and to control the use of pivoting and sampling. The input graph has to be connected.
- Betweenness Centrality: Constructs the Betweenness class for the given Graph G. If the betweenness scores should be normalized, then set normalized to True. The run() method takes $O(nm)$ time, where n is the number of nodes and m is the number of edges of the graph.
- Approximated Betweenness Centrality: Approximation of betweenness centrality according to algorithm described in Matteo Riondato and Evgenios M. Kornaropoulos: Fast Approximation of Betweenness Centrality through Sampling [9]. The algorithm approximates the betweenness of all vertices so that the scores are within an additive error ϵ (0.05 same as the approximated closeness) with probability $1-\delta$. The values are normalized by default. The run() method takes $O(m)$ time per sample, where m is the number of edges of the graph.
- Local Clustering Coefficient: Constructs the LocalClusteringCoefficient class for the given Graph G. If the local clustering coefficient values should be normalized, then set normalized to True. The graph may not contain self-loops. In fact at this point we have removed self-loops from the graph before the use of this method. This method was performed also with NetworkX, given the fact that it does not return an error for self-loops.
- Approximated Local Clustering Coefficient: Returns approximate average local clustering coefficient The maximum error can be given as a parameter and determines the number of samples taken.
- Global Clustering Coefficient: Returns the clustering coefficient of the graph.

Then, after the computation of these features for our original network, we did a comparison with random graphs using z-score and p-value. In particular Erdös-Renyi-Gilbert and Chung-Lu (same original graph characteristics in expectation) models are considered in this phase. We used methods from Networkit to generate such random graphs:

- Erdös-Renyi model: Creates random graphs in the G(n,p) model, where p is the probability of each edge to appear, independently of all other events.
- Chung-Lu model: Given an arbitrary degree sequence, the Chung-Lu generative model will produce a random graph with the same expected degree sequence.

The parameter p of the Erdös-Renyi model was chosen based on the: $\mathbb{E}[m(n,p)] = p\binom{n}{2}$, where $m(n,p)$ is the number of edges in the random graph generated with Erdös-Renyi model.

Z-score and p-value are computed using the definitions seen in class and following the Monte-Carlo Approach.

$$\frac{f - \mathrm{E}[X_\mathrm{F}]}{\sigma[X_\mathrm{F}]} \tag{1}$$

$$\begin{aligned} p(H_0) &= \mathrm{P}[X_\mathrm{F} \leq f | H_0 \text{is true}] \\ p(H_0) &= \mathrm{P}[X_\mathrm{F} \geq f | H_0 \text{is true}] \end{aligned} \tag{2}$$

We assume the Null Hypothesis that the feature measured in our network well conforms with the distribution observed in a random graph. In fact, we decided to create a specific number of random graphs for each model, and then we computed the considered feature on the random graph instances to compute z-score and p-value.

In particular the p-value measures the probability of observing a value for the feature considered at least as extreme as the value obtained in our network when the Null Hypothesis is true (in a random graph).

We are interested in understanding:

- if the value of our measured feature is higher than expected;
- if the value of our measured feature is lower than expected.

V. Results

It is interesting to take look at the degree distribution first of all. In fact, looking at Fig. 2, we can notice that in our network there are a lot of nodes (politicians) which have a fairly large number of connections, while we can see only a few number of politicians with a small degree. The 5 politicians with highest degree are reported on Table 2.

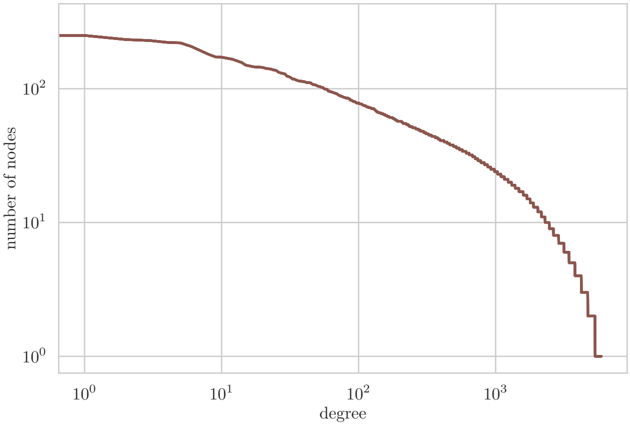


Fig. 2: Degree Distribution

7.1413in

Since nodes with a higher score of centrality are the pages of the politicians which are more central in the network, it is interesting to see that Sylviane Bulteau is returned with top centrality by Closeness centrality, Betweenness centrality, Approximated Closeness centrality and Approximated Betweenness centrality. So, we can say that Sylviane Bulteau is the most central politician in our network given the fact that all the centrality methods return her name in first position.

Politicians	Centrality
Sofie Carsten Nielsen	323.0
Alain Leboeuf	250.0
Sylviane Bulteau	233.0
Oliver Wittke	229.0
Zeca Dirceu	222.0

TABLE 2: Degree Centrality top 5

Politicians	Centrality
Sylviane Bulteau	0.358857
Assemblyman Ray Walter	0.323508
Norbert Spinrath	0.320471
Alexander Kulitz	0.317422
Alejandra Morlan	0.312467

TABLE 3: Closeness Centrality top 5

Politicians	Centrality
Sylviane Bulteau	0.268308
Sofie Carsten Nielsen	0.055521
Margaret Quirk MLA	0.054146
Pierre Moreau	0.048274
Reinhard Brandl	0.045088

TABLE 4: Betweenness Centrality top 5

Politicians	Centrality
Sylviane Bulteau	0.346021
Assemblyman Ray Walter	0.320513
Norbert Spinrath	0.316456
Alejandra Morlan	0.314465
Sofie Carsten Nielsen	0.306748

TABLE 5: Approx. Closeness Centrality top 5

Politicians	Centrality
Sylviane Bulteau	0.254962
Sofie Carsten Nielsen	0.058179
Margaret Quirk MLA	0.056468
Pierre Moreau	0.046886
Reinhard Brandl	0.045517

TABLE 6: Approx. Betweenness Centrality top 5

We decided to report only the top 5 politicians with highest centrality, but if we consider the top 10, we can also say that the four algorithms return a consistent result, since also the other politicians within the top centralities are more or less the same: for example Margaret Quirk MLA and Pierre Moreau are always in the top 10 with all the four centrality methods.
It is also important to remember that the top politicians can slightly change with the use of approximated methods.

In addition, looking to the clustering coefficient, it is possible to see that Sylviane Bulteau has a clustering coefficient around 0.66, and this means that the nodes (politicians) neighbors of Sylviane Bulteau tends to be clustered together with a quite high probability. It is not the politician with highest clustering coefficient, but anyway, it has an high value. In fact, for example, Pierre Moreau has a clustering coefficient of 0.75, which is higher. It means that for some politicians, even if they are not the most centrals ones, it is more probable that their neighbors are clustered together. A consistent check of that is given by the fact that the politicians with highest degree on Table 2 tend to be the ones with the top centrality values.

In order to validate this results we also performed a manual check by googling the results returned by the algorithms and indeed they seems to be important figures.

Both the Erdös-Renyi and Chung-Lu models can generate disconnected graphs, for this reason sometimes we can see betweenness equal to infinite or standard deviation equal to zero that returns an infinite z-score. In order to represent this values we clipped the z-score to the maximum non-infinite value, for this reason we can see a pronounced tail in the boundaries of the histograms. In some other cases we can have NaN values for z-score because of standard deviation equal to zero and with $f - \mathrm{E}[F] = 0$.

Then, in order to perform a significance check we used z-score and p-value to compare the results obtained from the real graph and the random ones to evaluate if the observations obtained in the real network are meaningful. In particular, the significance check for the Closeness Centrality will be reported in this chapter, while the same approach have been used for the other features and the plots can be find in the Appendix. It is important to remember that the results can slightly change since the two models for random graphs produce every time a graph completely at random among the possible graphs with the given characteristics. In Fig. 3 and Fig. 4, the frequencies of z-scores are reported for the closeness centrality under the Erdös Renyi and the Chung-Lu models.

We decided to plot also the violin plots of the z-scores in order to have a different view of the data, plotting the entire distribution. In fact, on Fig. 5 and Fig. 6, it is possible to see the range of the closeness centralities divided now in uniform blocks, and for each of them a violin representing the z-scores for each subset of the centrality is shown with the white dot in the middle of each violin which indicates the median. Instead, the black bar indicates the interquartile range an the black line defines the lower adjacent value and the upper adjacent value. All the other points are considered as outside points.

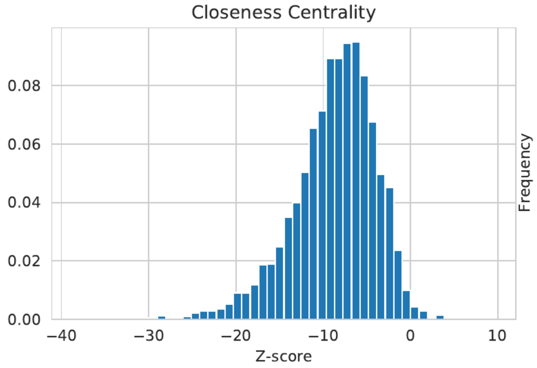


Fig. 3: Frequencies of z-scores with Erdös Renyi Closeness Centrality

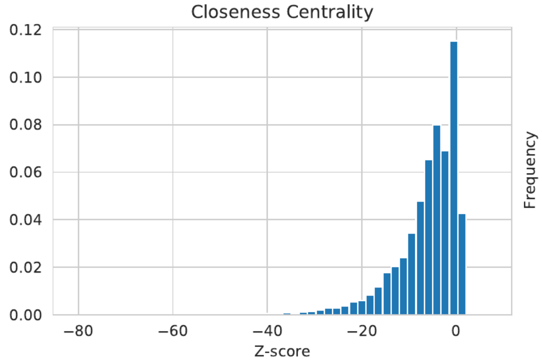


Fig. 4: Frequencies of z-scores with Chung-Lu Closeness Centrality

As we can see from Fig. 3 and Fig. 5, the z-scores considering the Erdös Renyi model tend to be in the range [0,-30], with a peak on -10, meaning that the random graphs produced by this model do not conform properly with our network. It means that the Null Hypothesis is rejected and our feature (closeness centrality) is significant in this case. Instead, as regards the Chung-Lu model, we have that the z-scores are mostly concentrated between -20 and 0, with a peak more close to 0

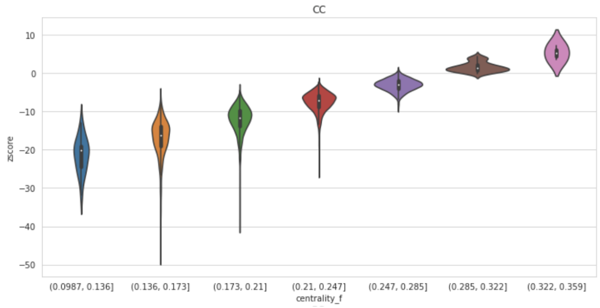


Fig. 5: Violin plot of z-scores with Erdös Renyi Closeness Centrality

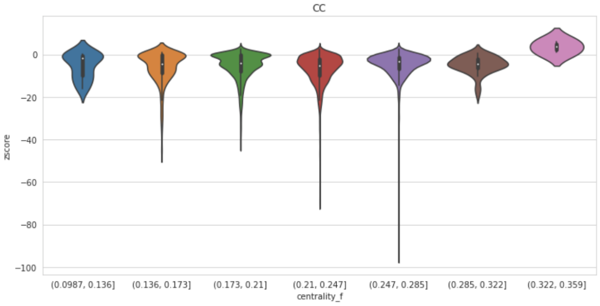


Fig. 6: Violin plot of z-scores with Chung-Lu Closeness Centrality

w.r.t. the Erdös Renyi model and this means that in this case the distributions observed on random graphs are more close to the one of our network. This is due to the fact that the graphs produced by the Chung-Lu model are more similar to real networks.

Considering p-value, we computed the two probabilities as defined before and we decided to print them for the top positions returned by our features of interest. We called them pvalue-ge (greater or equal) and pvalue-le (lower or equal). In general with the Erdös Renyi model considered pvalue-ge and pvalue-le tend to be around 0 and 1 repectively. While, considering the Chung-Lu model, they tend to assume different values between 0 and 1 depending on the case. As previously said, it is due to the fact that Chung-Lu model creates graphs more similar to real networks. As a final consideration, from Fig. 15 and Fig. 16 we can also notice that closeness centrality and betweenness centrality have different distributions of p-values with Chung-Lu model and it is possible to see from the p-values that there are some features that are significant for some politicians, while for some other politicians they are not significant because they well conforms with the distributions observed on random graphs.

VI. Conclusions

We found out that Sylviane Bulteau is the most central politician in our network, given the fact that in all the centrality measures she is the politician with highest score. Also the clustering coefficient returned an high value for her, as well as the high degree she has. Obviously there are other politicians with quite high values for the features considered. In addition, it is possible to see that politicians with high values of centrality tend to have also a high degree and a high clustering coefficient, meaning that they have a lot of neighbors and that they tend to be connected among each other. Finally, as a consequence of this, looking ad the significance check performed and according to our Null Hypothesis, we can confirm that not all the politicians have significant features.

References

[1] G. Mazzoleni and W. Schulz, “”mediatization” of politics: A challenge for democracy?,” *Political Communication*, vol. 16, pp. 247–261, 08 2010.

[2] H. Gil de Zuniga, N. Jung, and S. Valenzuela, “Social Media Use for News and Individuals’ Social Capital, Civic Engagement and Political Participation,” *Journal of Computer-Mediated Communication*, vol. 17, pp. 319–336, 04 2012.

[3] C. Fuchs, “The political economy of privacy on facebook,” *Television & New Media*, vol. 13, no. 2, pp. 139–159, 2012.

[4] D. R. Lisa M.Kruse and J. R.Flinchum, “Social media as a public sphere? politics on social media,” *The Sociological Quarterly*, vol. 59, no. 1, pp. 62–84, 2017.

[5] O. Serrat, *Social Network Analysis*, pp. 39–43. Singapore: Springer Singapore, 2017.

[6] R. A. Rossi and N. K. Ahmed, “The network data repository with interactive graph analytics and visualization,” in *AAAI*, 2015.

[7] B. Rozemberczki, R. Davies, R. Sarkar, and C. Sutton, “Gemsec: Graph embedding with self clustering,” in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2019*, pp. 65–72, ACM, 2019.

[8] E. Cohen, D. Delling, T. Pajor, and R. F. Werneck, “Computing classic closeness centrality, at scale,” *CoRR*, vol. abs/1409.0035, 2014.

[9] M. Riondato and E. M. Kornaropoulos, “Fast approximation of betweenness centrality through sampling,” *Data Mining and Knowledge Discovery*, pp. 438–475, 2016.

Appendix

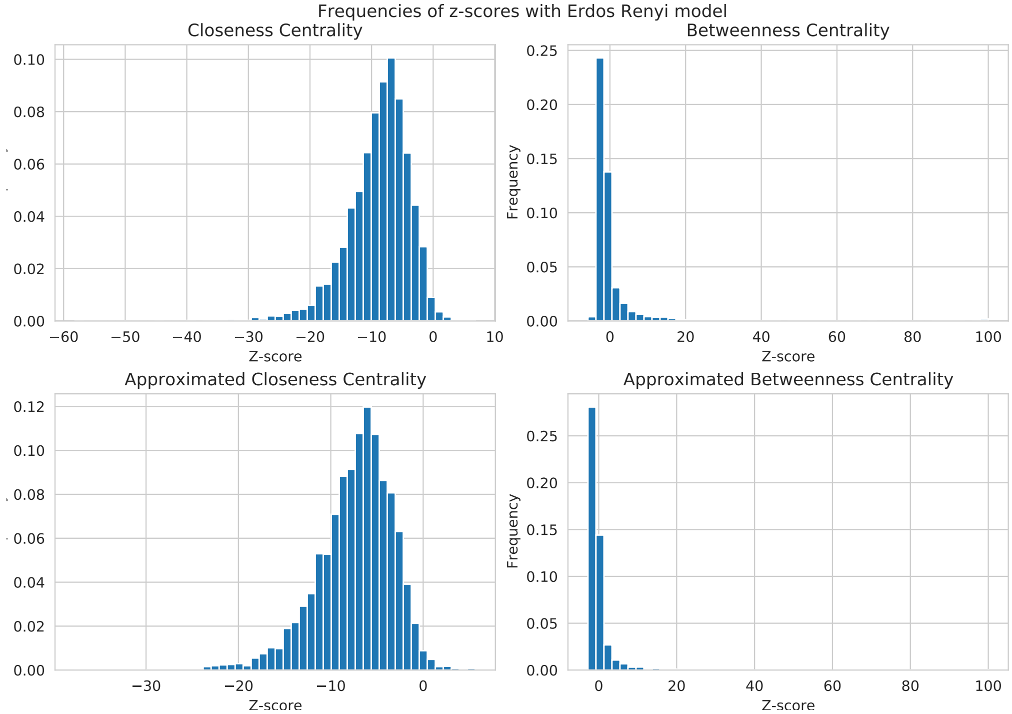


Fig. 7: Frequencies of z-scores with Erdos Renyi, Closeness Centrality, Betweenness Centrality, approximated Closeness Centrality, approximated Betweenness Centrality

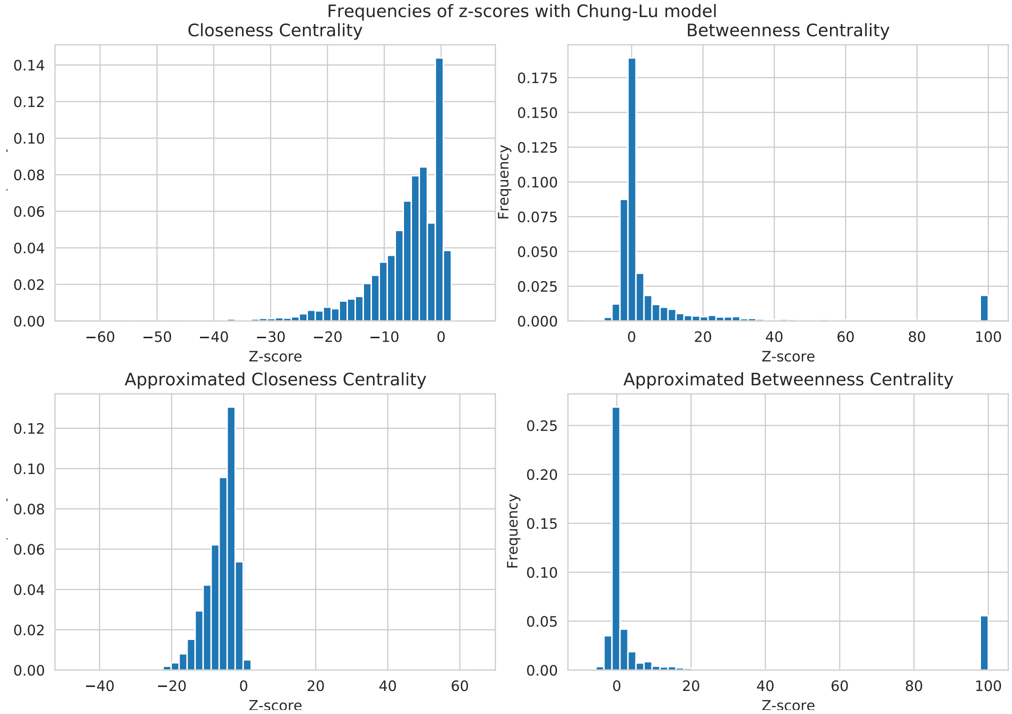


Fig. 8: Frequencies of z-scores with Chung-Lu

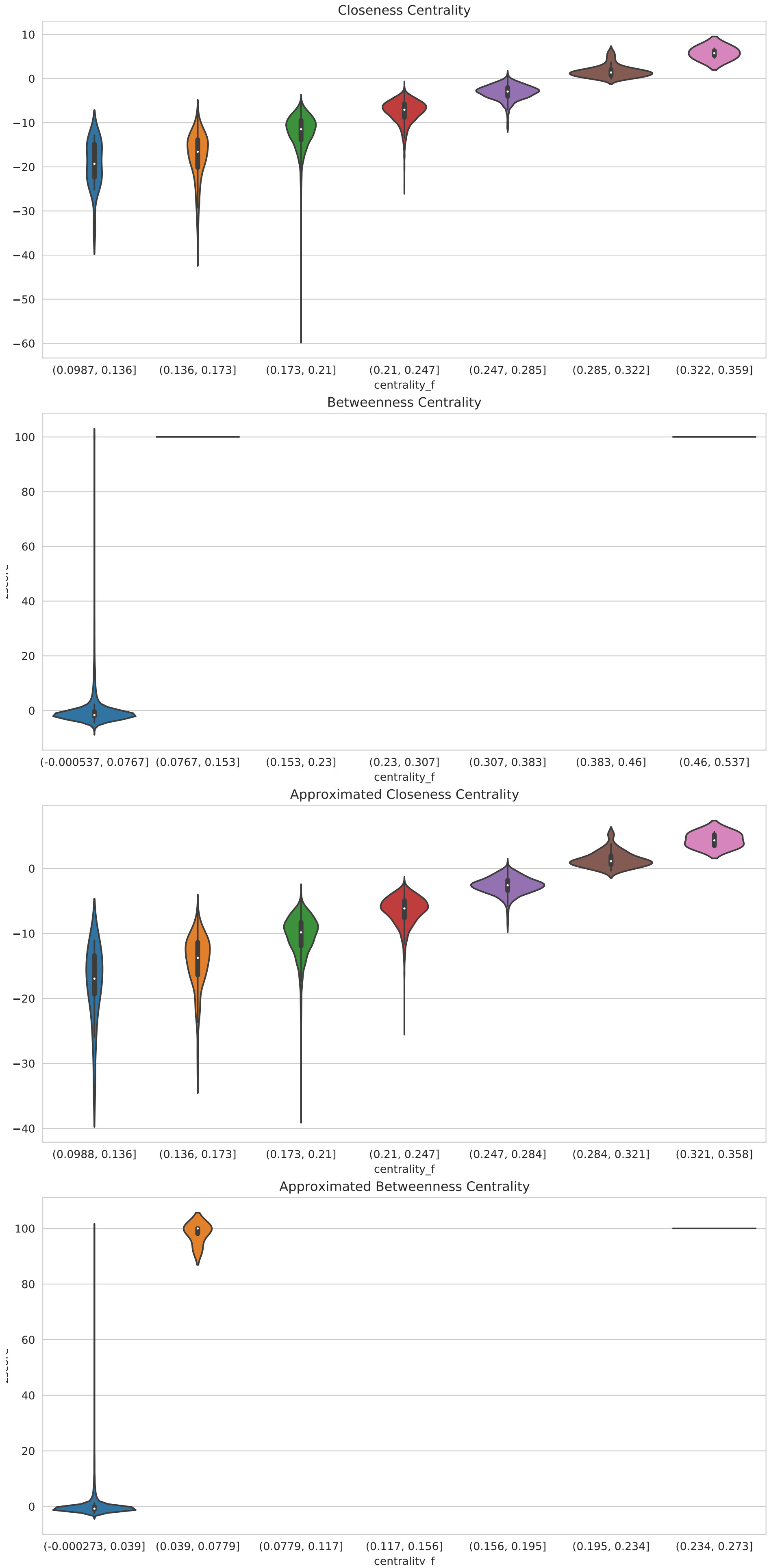
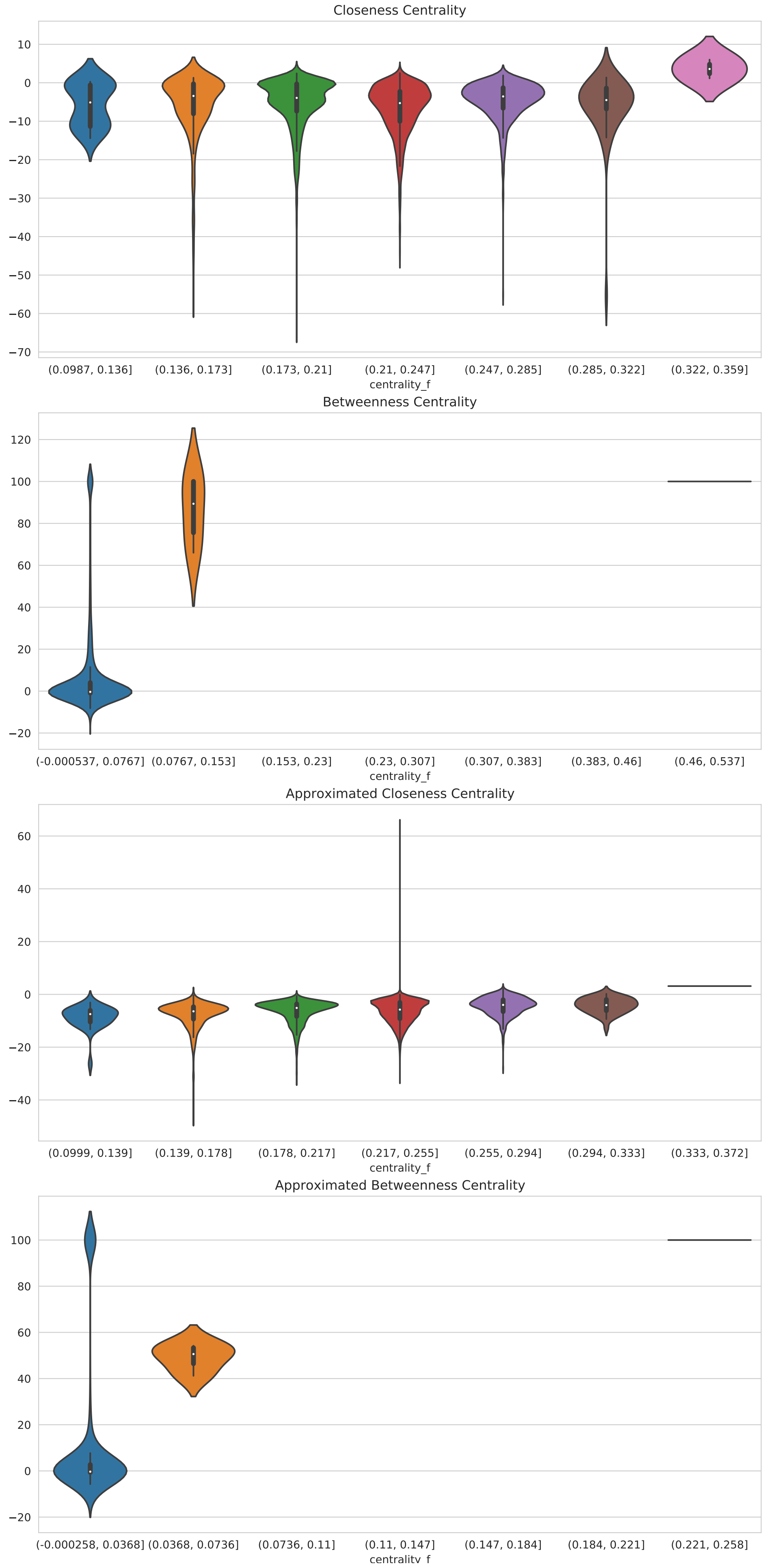


Fig. 9: Violin plots of z-scores with Erdos Renyi



Closeness Centrality

Betweenness Centrality

Approximated Closeness Centrality

Approximated Betweenness Centrality

Fig. 10: Violin plots of z-scores with Chung-Lu

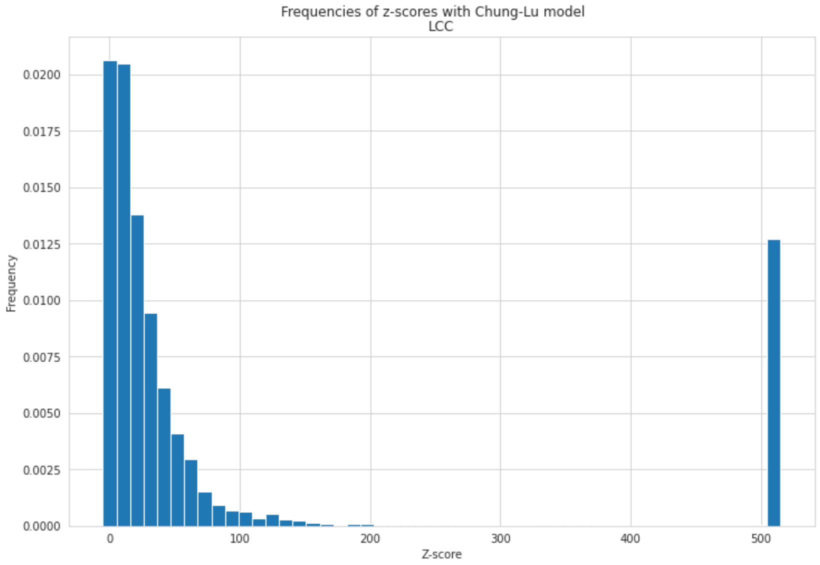


Fig. 11: Frequencies of z-scores with Chung-Lu for Local Clustering Coefficient

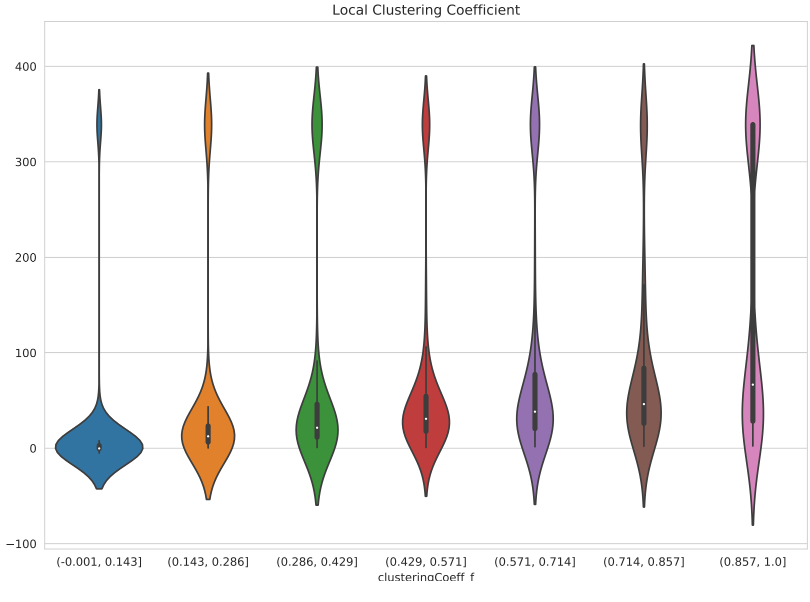


Fig. 12: Violin plots of z-scores with Chung-Lu for Local Clustering Coefficient

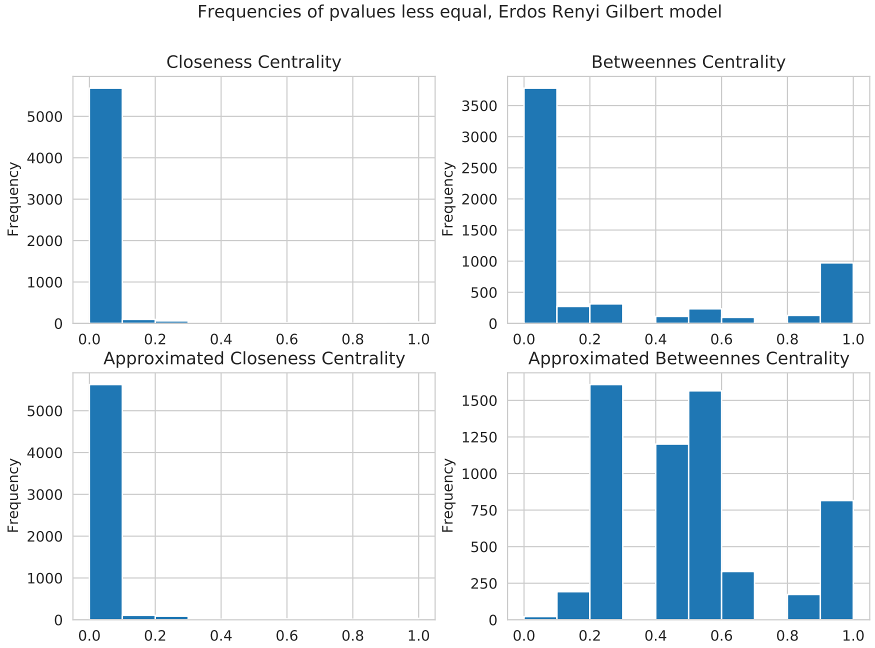


Fig. 13: Frequencies of p-values less equal, Erdos Renyi Gilbert model

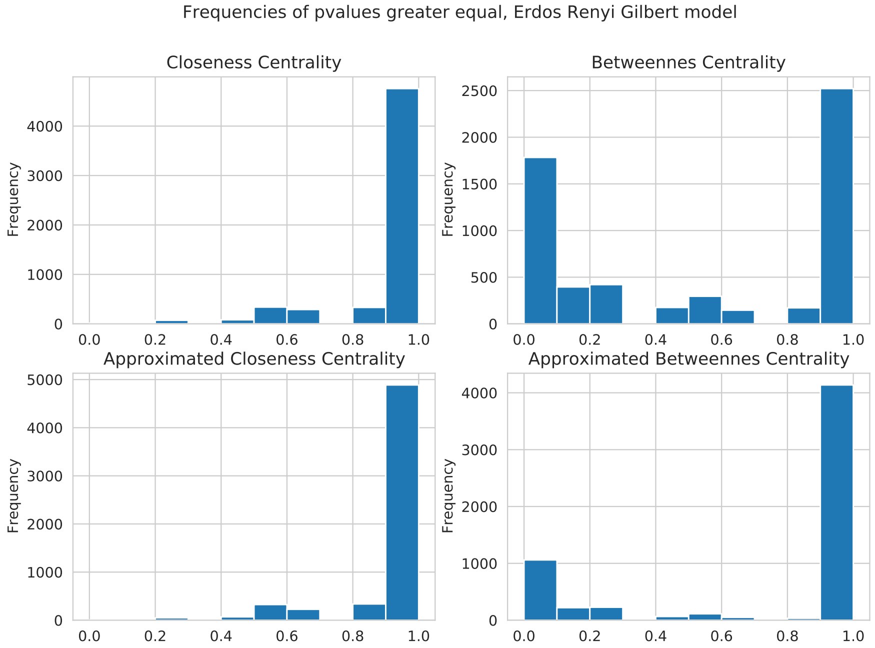


Fig. 14: Frequencies of p-values greater equal, Erdos Renyi Gilbert model

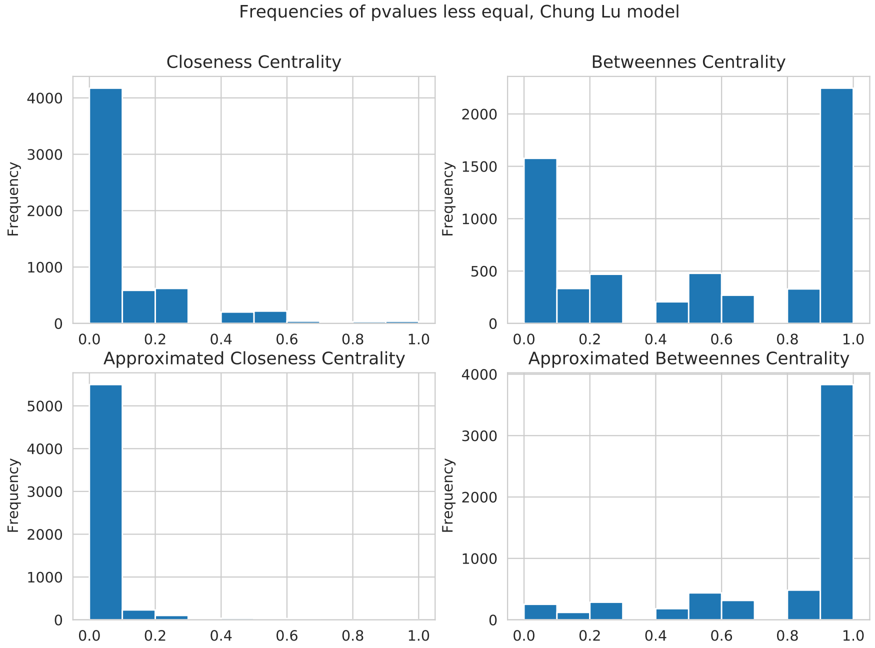


Fig. 15: Frequencies of p-values less equal, Chung Lu model

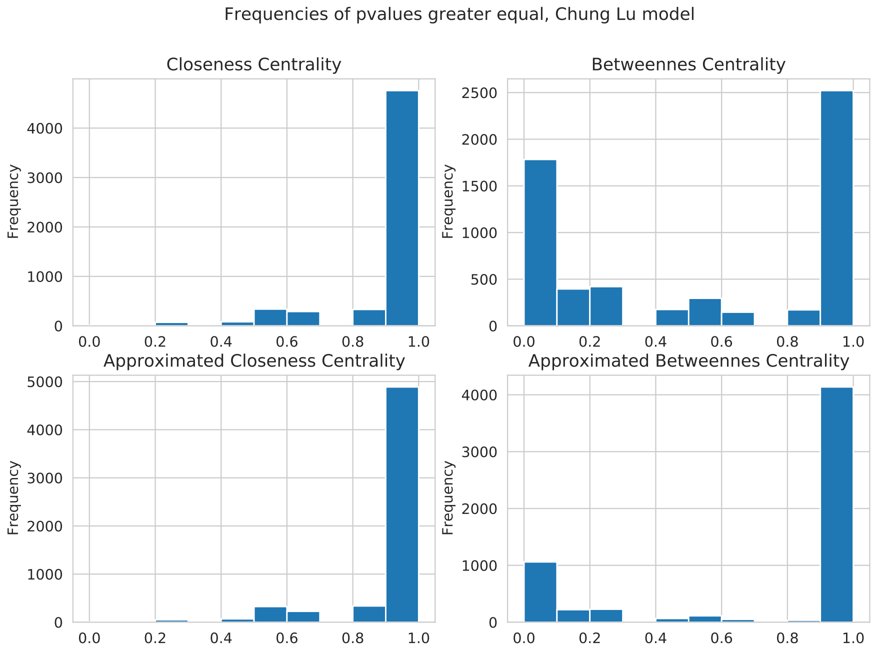


Fig. 16: Frequencies of p-values greater equal, Chung Lu model

CONTRIBUTIONS

A. Matteo Piva

- Initial analysis on the network to find if the graph is connected or disconnected and directed or undirected and its connected component(s), degree ditribution, diameter (exact and effective).
- Computation of global and local clustering coefficient.
- Generation of random graphs (Erdös-Renyi-Gilbert and Chung-Lu models used).
- Implementation of histogram and violin plots for z-score.
- Plot tests of the network with Gephi (due to the complexity of the network, the plot in Fig. 1 is very basic because Gephi crashed after a while).

B. Cristiano Colpo

- Initial import of the network as a graph.
- Analysis of centrality (closeness, betweenness, approx. closeness and approx. betweenness) and creation of dataframes with top centralities.
- Significance check with z-score and p-value at node level and graph level (global clustering coefficient) and creation of dataframes with z-score and p-value for each node.
- Use of Map-reduce paradigm when possible, the code can be easily modified with the library dask in order to run on multiple machines (the Monte Carlo phase).

C. Observations

The report has been done with LateX by Matteo and Cristiano in cooperation, taking the most important information from the notebook.

The workload has been divided with regular zoom meetings and fixing deadlines for each member.

Finally, as previously announced by email, due to a problem with the possibility to insert Learning from Networks in the study plan, differently from the previous two submissions, this one has been done only by two students: Matteo Piva and Cristiano Colpo.