Replication and impact analysis of "Scale-free topology of e-mail networks"

Jack Margeson*
University of Cincinnati, College of Engineering and Applied Science

(Dated: November 16, 2023)

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1 Introduction

This report serves as both a replication study and impact analysis on the text "Scale-free topology of e-mail networks" [1] by Holger Ebel, Lutz-Ingo Mielsch, and Stefan Bornholdt.

Following this introduction is a summary outlining the goals and results of the original paper. The summary leads into the replication of major findings from the paper, completed with a cleaned copy of the original dataset using NetworkX, a Python package with the purpose of creating, manipulating, and analyzing complex networks. The full source code for all replications performed can be found in a public GitHub repository [4] from the author of this report.

After the presentation of replications, an analysis is conducted on two identified examples of impactful work that have been produced citing the "Scale-free topology of e-mail networks" paper. Finally, a section theorizing two interesting directions for future research in the related field is included.

2 Summary

In the original text, Ebel et al. study a network consisting of sent and received e-mails. The data set used for the study was obtained by observing the log files of an e-mail server located at Kiel University. The authors of this paper were able to construct a network based upon this data, in which the nodes represent individual e-mail addresses and links between them represent a delivered e-mail. Several mathematical analyses were conducted in order to deduce information that could help classify the network.

The general conclusion found from these analyses was that their network " exhibits a scale-free link distribution and pronounced small-world behavior" [1] that aligns with the findings of similar studies on various other social-based networks. Implications for these findings, including the importance of the scale-free property of the network on the spread of e-mail viruses, are touched upon briefly towards the end of the paper.

2.1 Strengths

The greatest strength of the original paper is undoubtedly that it established a line of reasoning to determine the degree distribution of e-mail networks to exhibit a power law, providing evidence towards e-mail networks being scale-free. Additionally, not only did the authors find evidence that e-mail networks could be considered scale-free, they also found that the e-mail network that they had created from their sample data exhibited small world behavior, which indicates that node neighbors of a node are likely to be connected themselves. These two findings have been instrumental in future e-mail related studies, two of which will be talked about in the impact analysis section of this report.

*Additional author information: https://marg.es/on

2.2 Weaknesses

There are a few small albeit important weakness to mention present in the original paper. Justifications for these weakness are provided via opinion of the author of this report.

One issue that arises in the fact that the sampling process is restricted to exclusively one e-mail server. However, it would generally be unreasonable for the authors of the paper to acquire data from other universities or similar group-based entities with an emphasis on e-mail communication. This is due to the fact that a data set of e-mails as large as the one presented in the original paper would have to be extensively cleaned as to not expose sensitive data, which would take time and energy on the part of the group providing the dataset. Given this information, it is unlikely that another dataset to enforce preliminary findings would not be feasible to obtain for the original paper authors, leaving the task of replication with alternate datasets to other publications.

While not directly an issue with the paper itself, an problem arises with the dataset commonly distributed for this paper. As part of this replication research, a copy of the dataset used in the original paper was taken from the "Network Datasets" section of the site hosting the "Network Science" book by Albert-László Barabási [5]. However, this dataset is mildly incomplete when comparing some of the provided data to the stated statistics in the paper. For starters, a quick NetworkX network reconstruction and analysis of the graph shows us that there are a total of $N = 57,194$ nodes provided, while in the paper, it is stated that a total of $N = 59,912$ nodes exist overall, with a subset of $N = 56,969$ nodes making up a giant component in the network. Other inconsistencies, such as the lack of labels when it comes to which nodes are considered student e-mail nodes, has made it difficult to accurately replicate findings, which will be touched upon for each of the replications presented in this report.

3 Reproduction

As stated in the weaknesses section of this report, the distributed dataset has a few small problems when it comes to making a one-to-one replication of the original study. Below are the findings from the replication study. In cases where values or figures differ from the original paper, a justification is provided for why the discrepancy might have occurred. Findings that are unreproducable given the dataset are also discussed.

3.1 Network analysis

There are a few examples of network analysis that do not tie directly into a figure from the original paper. This section talks about these from a mathematical standpoint, without the use of figures. The first, which was mentioned previously, is the size of the network. Given the edgelist, we are able to reconstruct the graph structure through NetworkX to see that a total of $N = 57,194$ nodes exist in the graph. The paper states in a footnote that three email addresses have been excluded as artifacts, due to the fact that they represented such a large degree as to be designated an outlier. By sorting our list of nodes by degree, we are able to query the last five, resulting in the following list:

$$[('13498', 802), ('11798', 1020), ('13678', 1228), \\ ('11028', 4171), ('32199', 6553)]$$

We see that the last five nodes here have a degree greater than 800. Putting that in context of our e-mail scenario, these nodes are e-mail addresses that have connected to at least eight hundred other e-mail addresses (either receiving or sending an e-mail). This can be most likely explained by email servers, such as the university sending emails out to large subsets of students, as explained in the original paper. In order to reduce the impact that these nodes have on our distributions, we will remove them from any future calculations unless otherwise stated.

After the removal of the outlier nodes described, we sum the degrees of all nodes and divide by $N = 57,189$ to receive a mean degree of $< k >= 3.01$, which is close to the mean degree of nodes of the giant component listed in the original study, $< k_{large} >= 2.96$.

One issue with the dataset is that it is completely stripped of all data labels. This means for replication steps that require student-only nodes, we will have to approximate. Figure 2 in the original paper outlines the student account degree distribution, where only nodes that represent student e-mail addresses are contained. An approximation of this student-only dataset can be created by limiting the max degree that we consider over the network. We will define k_{upper} to represent the maximum degree that we are willing to include in our student-only node network. To find a reasonable value for k_{upper}, we can take a look at the mean degree of just the student nodes, identified in the paper as $< k >= 2.88$. Using this mean degree and a sorted list of nodes by their degree, we can remove n nodes from the distribution until we get a mean approximating $< k >= 2.88$:

$$summationequation$$

Using NetworkX, we find that removing the highest 27 entries from the dataset (22 more nodes removed than the 5 outliers removed previously), we find an average degree $< k >= 2.8793 \approx 2.88$, giving us $k_{upper} = 266$. Putting this number in context, it means that the maximum number of e-mail addresses that we'd reasonably expect a student e-mail address to be connected to is 266. We make the assumption that any of the nodes we've just removed were result of either smaller university owned e-mail accounts or students that receive an unusual amount of mail (teaching assistants, perhaps). Again, the reason that we must speculate the number of student nodes here is due to the fact that the dataset provided contains no labels on if an e-mail address node belongs to a student or not. The subset of nodes with degree equal to or less than 266 that we have just created will be used for calculating student e-mail account exclusive statistics and degree distributions.

3.2 Degree distribution

The first degree distribution, shown in the first figure [1], encompasses all nodes in the graph minus the outliers. In the network analysis section, we removed outliers from our supplied data set. While the exact parameters for graphing were not provided, we can assume that a bin size greater than one is used due to the fact that there is a limited number of points on the plot far less than the actual number of total nodes. Graphing this revised distribution on a double logarithmic scale with a bin size of 100 gives us a figure with similar features to that of the original:

3.3 Student account degree distribution

3.4 In/Out-degree distribution

4 Impact Analysis

[3][2]

5 New Directions

6 Conclusion

References

[1] Holger Ebel, Lutz-Ingo Mielsch, and Stefan Bornholdt. “Scale-free topology of e-mail networks”. In: *Phys. Rev. E* 66 (3 Sept. 2002), p. 035103. DOI: 10.1103/PhysRevE.66.035103. URL: https://link.aps.org/doi/10.1103/PhysRevE.66.035103.

[2] Ching-Hao Mao, Hahn-Ming Lee, and Che-Fu Yeh. “Adaptive e-mails intention finding system based on words social networks”. In: *Journal of Network and Computer Applications* 34.5 (2011). Dependable Multimedia Communications: Systems, Services, and Applications, pp. 1615–1622. ISSN: 1084-8045. DOI: https://doi.org/10.1016/j.jnca.2011.03.030. URL: https://www.sciencedirect.com/science/article/pii/S1084804511000841.

[3] Boli Xie and Maoxing Liu. “Dynamics Stability and Optimal Control of Virus Propagation Based on the E-Mail Network”. In: *IEEE ACCESS* 9 (2021), pp. 32449–32456. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3059767.

[4] Jack Margeson. *Email Topology Analysis.* Nov. 2023. URL: https://github.com/jack-margeson/email-topology.

[5] Albert-László Barabási. *Network Datasets.* URL: http://networksciencebook.com/translations/en/resources/data.html.