Course Project: Analysis of the ArXiv General Relativity and Quantum Cosmology Collaboration Network

Qiaoliang Xiang
qiaoliangxiang@gmail.com

1 Introduction

Social networking is one of the important parts of our daily life because it enables us to communicate, exchange information and collaborate with others. We are interested in studying the arXiv GR-QC (General Relativity and Quantum Cosmology) collaboration network [1], and we created a data collection of papers submitted between July 1992 and November 2013. A variety of analyses and visualizations have been performed on the collaboration network. It is a scale-free network whose node degrees follow a power-law. The sizes of connected components also follow a power-law, and the giant component contains a large amount of authors. It is a small world network with six degrees of separation. The network contains 3553 communities with modularity 0.85. We also studied how the graph evolved over years and found the network was becoming denser and its diameter was shrinking.

2 Data Collection

Jure Leskovec provides such a data collection that covers papers in the period from January 1993 to April 2003 (124 months), which is represented as an undirected and unweighted collaboration network [2]. His data collection, however, cannot satisfy the needs of our intended analysis. First, it cannot completely represent the GR-QC collaboration network since the papers submitted after April 2003 are not included. Second, we cannot effectively link our analysis to the real arXiv GR-QC network. Each author is represented as an integer and no name is given. If we find an author with the largest degree, we do not know who he is as well as validating whether he really has a large number of collaborators. Besides, no information of papers is available such as their titles, authors, and publication dates. Without the information of both authors and papers, it is not possible to assign weights to the collaboration network. To address the above-mentioned problems, we created a raw data collection by crawling the list of papers submitted between July 1992 and November 2013 (257 months) from arXiv. For each paper, we extracted its URL, its title and its authors. However, there are some limitations of the raw data collection such as authors names cannot be correctly mapped to real authors. We applied a few cleaning steps to solve these problems.

2.1 Raw Data (HTML)

A paper on arXiv is organized according to the academic fields it belongs to and the time (year and month) when it is submitted. The information of the papers submitted in a given year and a given month is displayed on a web page, the URL of which can be constructed using a method that will

be described in the next paragraph. There are 257 months between July 1992 and November 2013. Consequently, we obtained 257 HTML files using Python [3] on 21 November, 2013. The file name of an HTML file is a combination of year and month. Note that the papers submitted after that day are not included.

All papers related to general relativity and quantum cosmology that were submitted in November 2013 can be accessed by visiting `http://arxiv.org/list/gr-qc/1311?show=1000`, where *qr-gc* means general relativity and quantum cosmology, *1311* means November 2013, and *show=1000* requires at most 1000 papers to be returned. Without using the parameter *show=1000*, only the first 25 papers are returned. That is why a large number 1000 was used to make sure all the papers submitted within a month can be returned.

2.2 Clean Data (XML)

We used Python again to create clean data from the raw data. Our clean data collection contains 257 XML files, each of which was generated from an HTML file. An XML element represents a paper, and the following information is stored for the paper: the year when it is submitted, the month when it is submitted, the URL of its abstract, its title, and the authors who wrote the paper. For instance, the XML element of the first paper submitted in July 1992 is displayed below.

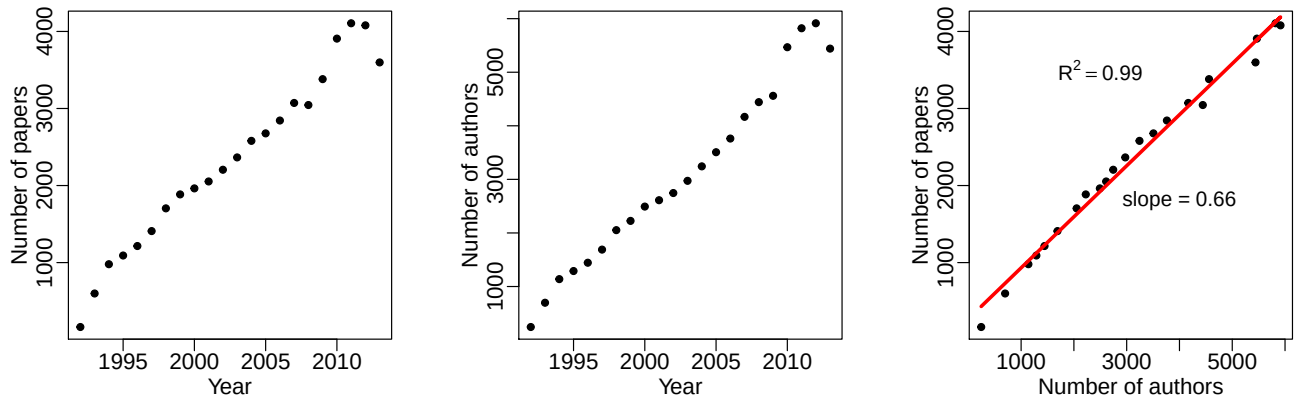
```
<paper year="1992" month="7" url="http://arxiv.org/abs/gr-qc/9207002">
  <title>Remarks on Pure Spin Connection Formulations of Gravity</title>
  <authors count="2">
    <author>riccardo capovilla</author>
    <author>ted jacobson</author>
  </authors>
</paper>
```

The HTML files are well structured so it is relatively easy to extract the URL, title and author names for each paper. The information of year and month is obtained from the name of an HTML file. The challenge part is how to associate author names to real authors. First, different authors may have the same name, resulting in an underestimate of the true number of authors. Second, an author's name in different papers may be written in different ways, leading to an overestimate of the true number of authors [4]. Based on the limited information, the first problem cannot be simply solved. Instead, we focus on solving the second problem by implementing many heuristic rules found using a trial-and-error approach.

We performed a few cleaning steps so as to generate a clean XML data collection. First, we applied the following transformations to each character of an author name.

1. Change the character to lower case. An author may have used upper case names, lower case names, and mixed case names. For instance, we found an author's name was written in two manners: *Kei–ichi Maeda* and *Kei-ichi MAEDA*. Normalizing names to lower case enables us to group different variants of an author's name together.

2. If the character is a dash (-) or an underscore (_), replace it by a space. An author can use dashes, underscores or spaces between different parts of his name. For example, we found an author's name was written in two different ways: *hector vargas-rodriguez* and *hector vargas*



(a) (b) (c)

Figure 1: (a) The number of papers per year. (b) The number of authors per year. (c) The number of papers per year versus the number of authors per year. The red line represents a fitted simple linear regression model with slope 0.66 and coefficient of determination 0.99.

rodriguez. Replacing the dash by a space in the former name helps us to relate them to the same author.

3. If the character has a diacritic mark, replace it with the corresponding English letter. We wrote programs to find those special characters within the data collection, and found the correct letters to replace them. One such rule is to use e to replace $\acute{e}, \ddot{e}, \grave{e}, \hat{e}$, and $\breve{e}$. The two different names *B. Gouteraux* and *B. Goutéraux* in the data collection will be considered to be the same author with the help of this step.

4. Delete the character if it is not one of these: a letter or a comma or a space or a dot. We found the following invalid characters in our collected data set: left curly bracket ({), right curly bracket (}), backslash (\), tilde (~), grave accent (`), caret (^), apostrophe (´), and semicolon (;). In the data collection, there are two names *Agnieszka Szyp lowska* and *Agnieszka Szyp\lowska*. After removing invalid characters, they can be considered to be the same author.

Second, we deleted duplicate spaces in every name so that author names are invariant to extra spaces. Third, duplicate authors within each paper were eliminated because we found a few papers contained duplicate authors by mistake. Fourth, we removed duplicate papers since we found a few papers were uploaded again. Each paper is assigned with a unique identifier that consists of its title in lowercase and the set of authors. Two papers are considered as the same if their unique identifiers are the same.

Table 1 summarizes a few key statistics of the raw data and the clean data. After the cleaning step, we have collected 50922 papers written by 27330 authors. There are 114554 collaborations between authors. Note that the data collection provided by Jure Leskovec has 5242 authors and 28980 collaborations [2]. The number of papers submitted per year and the number of authors who submitted at least a paper per year are shown in Figure 1(a) and Figure 1(b), respectively. The two figures are quite similar, suggesting that the two quantities could be strongly correlated.

We performed a linear regression [1] using the number of authors per year to predict the number of papers per year with the help of the R programming language [5]. Figure 1(c) shows the relationship between them. The coefficient of determination (R^2) [6] is 0.99, indicating the regression line perfectly fits the data.

Statistics	Raw Data	Clean Data
Number of Authors	28071	27330
Number of Papers	50967	50922
Number of Collaborations	129429	128125

Table 1: The key statistics of the raw data and the clean data.

3 Data Analysis

We performed data analysis and visualization using *R* [5] and Gephi [7]. We also provided all the necessary information so that others can replicate our methods. The major R functions are included in the footnote. Based on the clean data, we can construct two networks: authorship and collaboration [8]. An authorship network is a bipartite undirected graph that consists of both authors and papers as nodes with edges connecting every paper to its authors. A collaboration network is an undirected graph of authors where two authors are connected if they have coauthored a paper. The collaboration network is the focus of this study, and edges are unweighted unless otherwise stated.

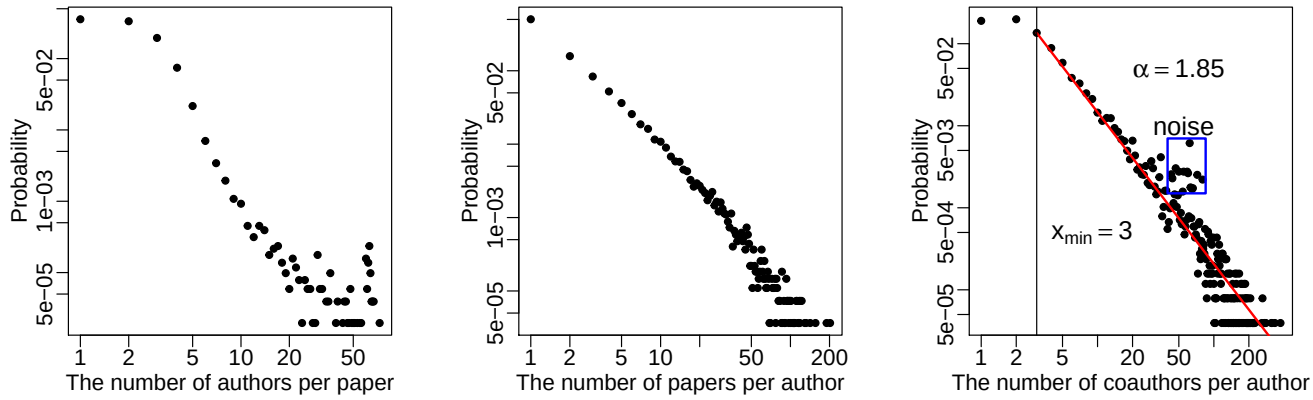
3.1 Degree Distribution

Number of Authors per Paper The number of authors per paper varies from 1 to 73. In average, a paper has only 2.25 authors. Figure 2(a) shows the distribution of the number of authors per paper on doubly logarithmic axes. The percentages and the cumulative percentages of papers with at most five authors are listed in Table 2. 35.6% papers were written by one author, and 33.4% papers were written by two authors. A large amount of papers (98.1%) has at most five authors. The paper with degree 73 is *Scientific Potential of Einstein Telescope* that was published in August 2011. The paper introduces the overall aims and objectives of the Einstein gravitational-wave Telescope (ET) and discusses its potential to influence our understanding of fundamental physics, astrophysics and cosmology [9].

Number of Authors per Paper	1	2	3	4	5
Percentage	35.6%	33.4%	19.5%	7.4%	2.2%
Cumulative Percentage	35.6%	69.0%	88.5%	95.9%	98.1%

Table 2: The percentages and cumulative percentages of papers having at most five authors.

[1] *lm* `http://stat.ethz.ch/R-manual/R-patched/library/stats/html/lm.html`.



(a) (b) (c)

Figure 2: (a) The distribution of the number of authors per paper on a log-log space. (b) The distribution of the number of papers per author on a log-log space. (c) The distribution of the number of collaborators per author on a log-log space. The vertical line marks $x_{min} = 3$, and the red line represents a fitted power-law distribution with $\alpha = 1.85$. The blue rectangle highlights the noise that may lead to an underestimate of the exponent.

Number of Papers per Author The number of papers per author falls into the range between 1 and 202. An author produced 4.19 papers in average. The distribution of the number of papers per author is plotted on doubly logarithmic axes in Figure 2(b). Table 3 lists the percentages and cumulative percentages of the authors who wrote less than 10 papers. 50.3% authors wrote only one paper, and 90.5% authors wrote less than 10 papers. 97% authors wrote less than 24 papers. The author with the largest number of papers (202) is *Lorenzo Iorio* who is an Italian physicist active in the field of general relativity and gravitation physics [10]. The author who published the second largest number of papers (190) is *Matt Visser*, whose research interests include black holes, general relativity, and cosmology [11].

Papers per Author	1	2	3	4	5	6	7	8	9
Percentage	50.3%	15.8%	8.3%	5.2%	3.6%	2.6%	1.9%	1.6%	1.2%
Cumulative Percentage	50.3%	66.1%	74.4%	79.6%	83.2%	85.8%	87.7%	89.3%	90.5%

Table 3: The percentages and cumulative percentages of authors who wrote less than 10 papers.

Power-Law Distribution A quantity x obeys a power-law distribution if its probability density function is $p(x) = Cx^{\alpha}$, where C is the normalization constant that ensures the sum of the probabilities over all x is one, and α is the exponent or scaling constant that usually takes values between 2 and 3 [12]. In practice, most empirical distributions follow power laws only for values greater than some value x_{min}, and the tail of the distribution is said to follow a power law. We used the code provided by Laurent Dubroca [13] to estimate x_{min} and α.

Number of Collaborators per Author The number of collaborators per author lies [2] between 0 and 374. In average, an author has 9.38 collaborators with a standard deviation of 19.8 collaborators. The percentages and cumulative percentages of authors having less than 9 collaborators are listed in Table 4. 7.1% authors have no collaborators, 17.7% authors have only one collaborator, and 55.8% authors collaborated with at most three scientists. Figure 2(c) shows the distribution of the number of collaborators per author and the fitted power-law distribution [3] with $x_{min} = 3$ and exponent $\alpha = 1.85$. The exponent is less than 2, which could be explained by the noise that is highlighted by the blue rectangle. If we remove all the authors who have more than 50 collaborators, the resulting exponent becomes $\alpha = 2.03$. Therefore, we can still conclude that the collaboration network is a scale-free network.

Number of Collaborators	0	1	2	3	4	5	6	7	8
Percentage	7.1%	17.7%	18.4%	12.6%	8.2%	5.4%	3.6%	3.0%	2.3%
Cumulative Percentage	7.1%	24.8%	43.2%	55.8%	64.0%	69.4%	73.0%	76.0%	78.3%

Table 4: The percentages and cumulative percentages of authors with less than 9 collaborators.

3.2 Connected Components

A connected component of an undirected graph is a subgraph in which any two nodes are connected to each other by paths. If a graph is connected, there is exactly one connected component. Otherwise, the graph has at least two connected components. The breadth-first search is usually used to find the connected components of a graph [14]. The size of a connected component is the number of nodes within it. The largest component is a giant component if it contains a significant fraction of nodes.

We identified 3446 connected components [4]. Every component is associated with a rank according to its size. The components with ranks larger than three follow a power law with exponent $\alpha = 3.13$ [15]. The giant component contains 74.2% authors, while the second largest component contains only 0.4% authors. There are 1940 connected components whose sizes are all ones, which correspond to the 7.1% authors without any collaboration. Figure 3(a) plots the size of a connected component versus its rank on doubly logarithmic axes. The frequency distribution of component sizes on doubly logarithmic axes is shown in Figure 3(b). The collaboration network is overall well connected.

3.3 Small-World Network

A small-world network is a type of network where most nodes can be reached from one another by a relative smaller number of paths. Such a network usually has a small average shortest path length and a much higher clustering coefficient [16]. The typical shortest path length between two randomly chosen nodes grows proportionally to the logarithm of the number of nodes. A related phenomenon is the six degrees of separation, which states that any two persons can be reached within six paths [17].

[2] *degree* `http://igraph.sourceforge.net/doc/R/degree.html`.
[3] *plfit* `http://tuvalu.santafe.edu/~aaronc/powerlaws/plfit.r`.
[4] *clusters* `http://igraph.sourceforge.net/doc/R/clusters.html`

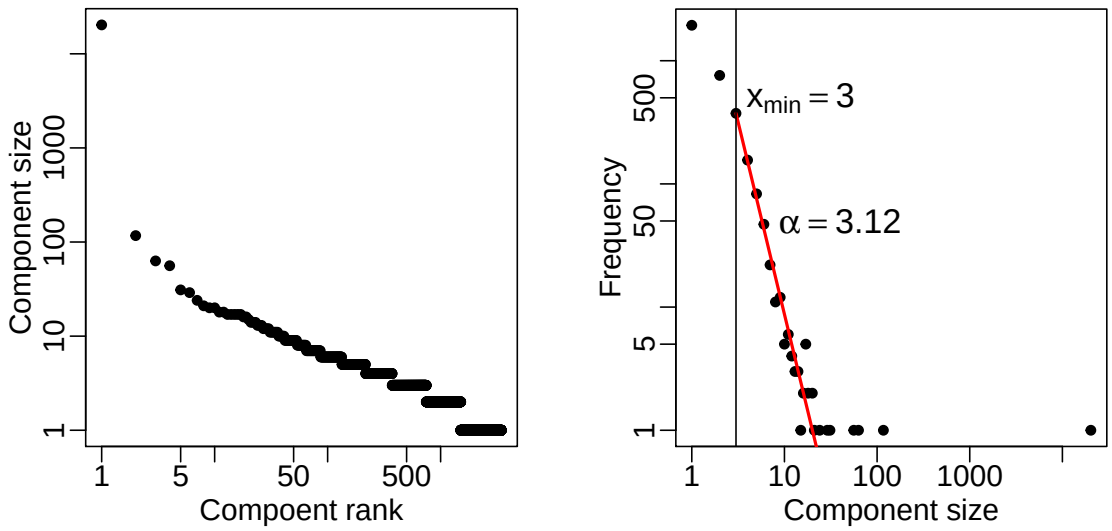
A clustering coefficient indicates how nodes are clustered. The global clustering coefficient is the ratio of the number of closed triplets to the number of triplets (both open and closed), where a triplet is formed by three nodes that are connected by either two (open triplet) or three (closed triplet) undirected edges. In an undirected graph, the local clustering coefficient of a node is the number of links between its neighbor nodes over the number of maximally possible links between them, and the local clustering coefficient of the graph is the average local clustering coefficients of all nodes with degrees larger than one [18].

We generated an Erdős - Rényi random graph [5] with the same numbers of nodes and edges as the collaboration network [19]. We compare them in a few aspects in Table 5. The six degrees of separation is valid in this collaboration network because the average shortest path length [6] is 6.15. The corresponding random graph that constructed by randomly connecting nodes would have an average distance $\frac{\log 27330}{\log 9.38} = 4.56$, where 277330 is the number of authors and 9.38 is the average degree of the collaboration network. The global clustering coefficient [7] (0.60) of the collaboration network is significantly higher than the global clustering coefficient (3.5×10^{-4}) of the corresponding random network. If two scientists collaborate with a third scientist, the probability that they will collaborate is 0.6. Therefore, we conclude that the collaboration network is a small-world network.

3.4 Community

Clustering is the partition of a set of nodes into mutual disjoint subsets (clusters) such that nodes within the same cluster are more similar to each other than to the nodes within the rest clusters.

[5] *erdos.renyi.game* `http://igraph.sourceforge.net/doc/R/erdos.renyi.game.html`.
[6] *average.path.length* `http://igraph.sourceforge.net/doc/R/shortest.paths.html`.
[7] *transitivity* `http://igraph.sourceforge.net/doc/R/transitivity.html`.



(a) (b)

Figure 3: (a) The size of a connected component versus its rank on doubly logarithmic axes. (b) The frequency distribution of component sizes on doubly logarithmic axes. The vertical line is placed at $x_{min} = 3$, and the red line represents the estimated power-law distribution with $\alpha = 3.12$.

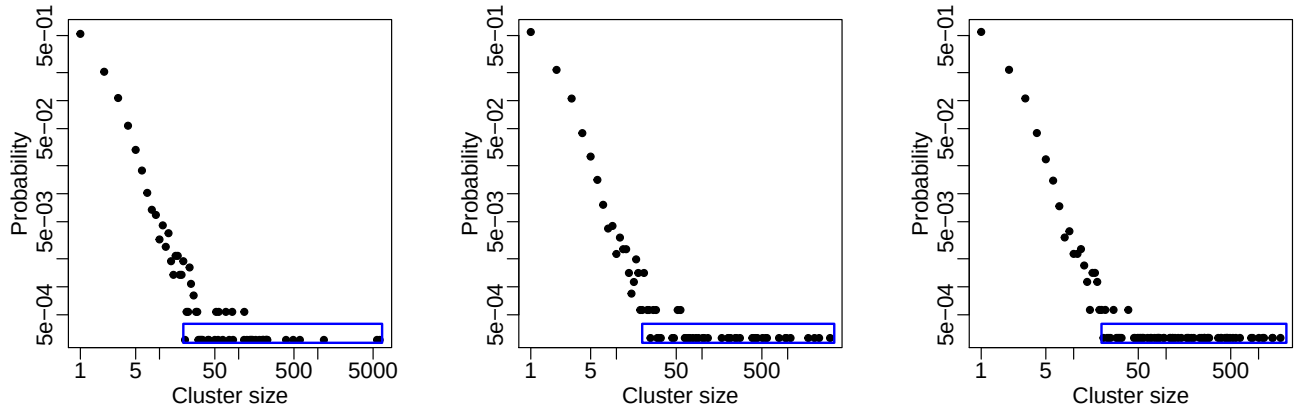
Network	ASPL	GCC	LCC
Collaboration Network	6.15	0.60	0.74
Erdős - Rényi Random Graph	4.56	3.5×10^{-4}	3.7×10^{-4}

Table 5: Comparisons between the collaboration network and its corresponding Erdős - Rényi random network in terms of the average shortest path length (ASPL), global clustering coefficient (GCC), and local clustering coefficient (LCC).

The quality of a clustering can be measured by a quality function called modularity, which is the difference between the fraction of within-cluster edges in a network and the fraction of the expected within-cluster edges in a random graph with the same node degree distribution as the given network [20]. A network with modularity larger than 0.3 usually contains significant community structure. Modularity is often used in optimization methods for detecting community structure in networks.

We want to compare three different weighting methods. Let δ_i^k be one if the i-th scientist is a coauthor of the k-th paper and zero otherwise. Denote n_k by the number of coauthors of the k-th paper. The weight between the i-th scientist and the j-th scientist is denoted by w_{ij}. The first weighting method is used in the unweighted collaboration network where the weight $w_{ij} = \max_k \delta_i^k \delta_j^k$ between two scientists is one if they have coauthored at least one paper and zero otherwise. The second weighting scheme takes the number of papers into account and defines the weight $w_{ij} = \sum_k \delta_i^k \delta_j^k$ as the number of papers coauthored between two scientists. The social bound between two scientists who coauthored a paper is stronger if the paper has less number of coauthors. To account for this effect, the third weighting method $w_{ij} = \sum_k \frac{\delta_i^k \delta_j^k}{n_k - 1}$ assigns a weight to each paper that is inversely related to the number of coauthors [21].

We used a fast greedy modularity optimization algorithm for finding community structure. The clustering discovered by the first weighting method contains 3721 clusters with modularity 0.79. The second weighting method produced 3553 clusters with modularity 0.82. A clustering with 3553 clusters and modularity 0.85 was founded by the third weighting method. The sizes of the top



(a) The first weighting method (b) The second weighting method (c) The third weighting method

Figure 4: The distributions of cluster sizes under the first weighting method (a), the second weighting method (b), and the third weighting method (c) on doubly logarithmic axes, respectively. The major differences in them are highlighted using blue rectangles.

Cluster Rank	1	2	3	4	5	6	7	8	9	10
The First Weighting Method	6051	5606	1195	594	489	397	230	229	222	221
The Second Weighting Method	3129	2331	2057	1740	1081	966	944	793	573	556
The Third Weighting Method	1727	1423	1411	1145	1110	1008	863	832	629	568

Table 6: The sizes of the top 10 largest clusters discovered using the three weighting schemes.

10 clusters discovered under the three weighting methods are listed in Table 6. We find the sizes of the largest cluster decreased almost by a half when a better weighting scheme is used. The distributions of cluster sizes under the three weighting schemes are plotted on doubly logarithmic axes in Figure 4. The highlighted blue rectangles in the three subfigures reveal that the cluster sizes are more evenly distributed with the use of a better weighting method. In summary, when a weighting scheme better reflects the social bounds between scientists, the clustering with higher modularity can be discovered, the sizes of the top clusters decrease, and the distribution of the cluster sizes is less skewed.

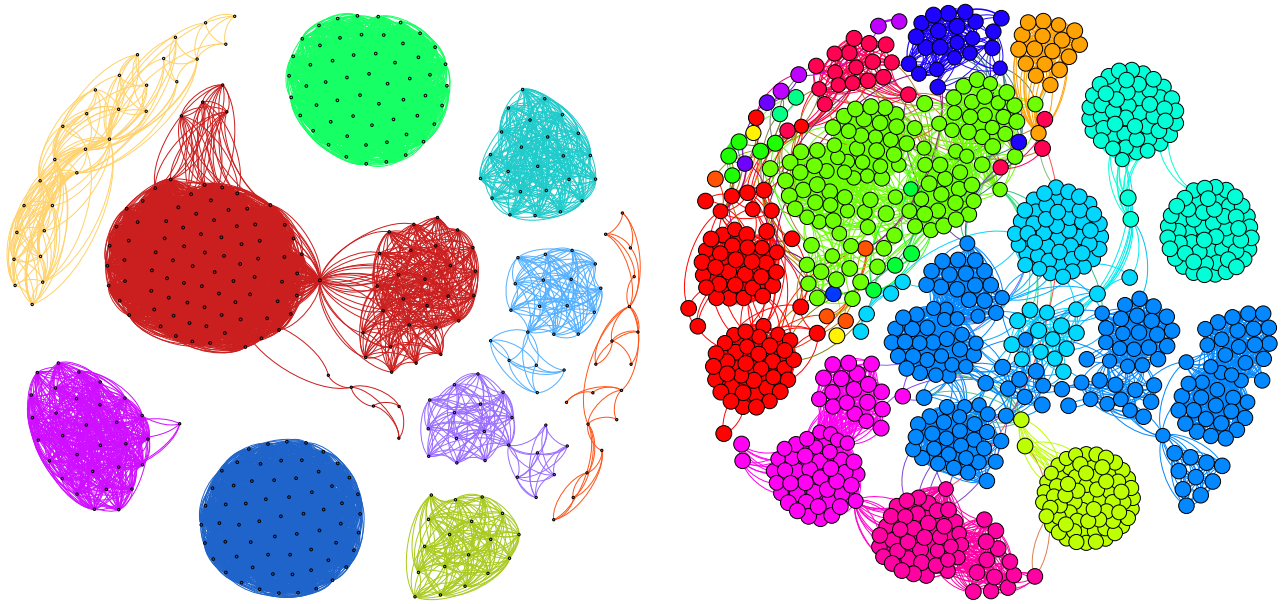
Visualization is important to help us understand the community structures in a visual way. We tried many kinds of visualizations and realized it's a challenge to visualize the network to convey essential and detailed structure information due to the large number of node and edges. We tried our best to visualize the unweighted collaboration graph. Two steps are taken to overcome the problem caused by the giant component that occupies a significant portion of the graph. First, we want to show each of the non-giant component is highly clustered by plotting the top 10 non-giant connected components in Figure 5(a) using the Fruchterman Reingold layout. Second, we focused on visualizing the communities within the giant component. We obtained a subgraph of 821 (3.23%) nodes and 13541 (10.57%) edges by selecting the nodes in the giant component whose degrees are between 50 and 100. The communities of the subgraph are visualized in Figure 5(b) using Fruchterman Reingold layout. It is obvious that each community is highly dense and structured, and most communities are connected by a few edges.

3.5 Graph Evolution

The collaboration network is evolving over time. In this Section, we studied how a few properties change over years. We created a series of graphs consisting of nodes and edges up to each year, and then performed various measures on graphs.

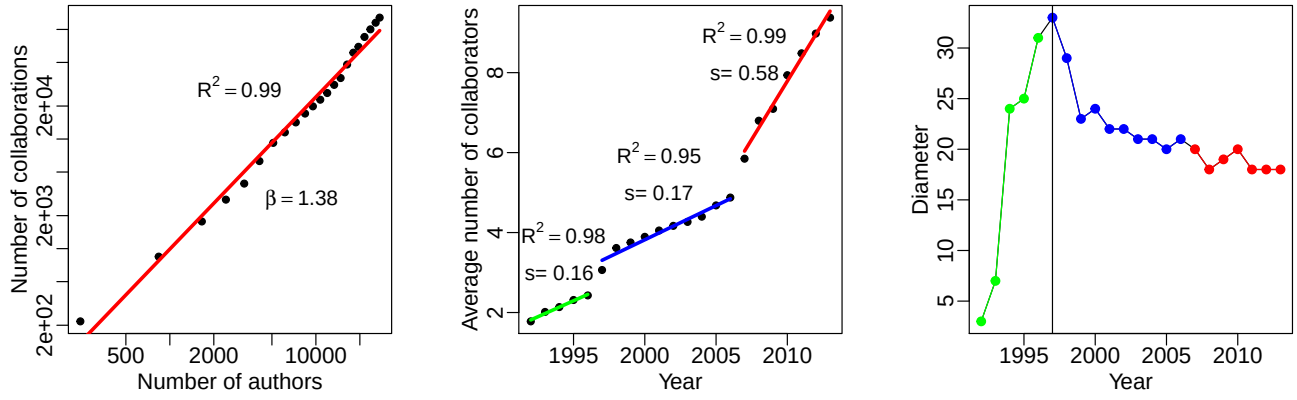
Densification Power Law Let $e(t)$ and $v(t)$ be the numbers of edges and nodes at time t, respectively. The densfication power law states they are related as $e(t) \propto v(t)^{\beta}$, where β is an exponent that generally lies between 1 and 2 [23]. Exponent $\beta = 1$ corresponds to constant average degree over time, while $\beta = 2$ corresponds to an extremely dense graph. The number of collaborations at each year versus the number of authors at that year is plotted on a log-log plot in Figure 6(a). We used a linear regression model to fit the data [8] in the log-log space. The slope of the line is actually the exponent $\beta = 1.38$ of the densification power law. The coefficient of determination is as high as 0.99, indicating a perfectly linear fit of the data in the log-log space.

[8] *lm* `http://stat.ethz.ch/R-manual/R-patched/library/stats/html/lm.html`.



(a) (b)

Figure 5: The two figures are plotted using the Fruchterman Reingold layout. (a) The top 10 non-giant connected components. (b) The communities of the subgraph whose nodes are the nodes in the giant component with degrees are between 50 and 100.



(a) (b) (c)

Figure 6: (a) The number of collaborations per author at a year versus the number of authors at that year on a log-log plot. The red line represents the estimated densification power law in the log-log space with slope or exponent 1.38 and coefficient of determination 0.99. (b) The average number of collaborators per author over years. (c) The diameter over years. The vertical line marks the gelling point.

Increasing Average Number of Collaborators per Author Figure 6(b) shows the average number of collaborators per author at each year. We identified three different stages of linear grows in the figure. The 22 years were divided into three ranges, and linear regression was applied to each range. The coefficients of determination of the three stages are very high. The average number of collaborators linearly increased at a speed of 0.16 in the first five years (1992-1996), and each author had an average number of 2.13 collaborators. In the next 10 years (1997-2006), it increased linearly at a slightly faster speed of 0.17 with each author having an average number of 4.08 collaborators. In the last 7 years (2007-2013), there was a dramatic increase in its speed that is as high as 0.58, and each author had an average number of 7.79 collaborators. Based on the above analysis, we conclude that the collaboration network is becoming denser over time and the number of collaborations grows super-linearly in the number of authors.

Shrinking Diameters The phenomenon of shrinking diameters refers to the observation that the effective diameter or diameter decreases as the network grows [24]. Even the effective diameter is more robust than diameter, it was shown that they exhibit qualitatively similar behaviors. Hence, we used diameter [9] for convenience. Figure 6(c) shows the diameter at each year, which is colored according to the three stages identified in the above paragraph. The network in the first stage (1992-1996) was *establishing* such that the diameter increased very quickly and contained many small and disconnected components. In the second stage (1997-2006), the diameter was *shrinking*. Year 1997 is called a gelling point where the diameter spiked and the largest connected component was formed by merging several small disconnected components [15]. After the gelling point, the inclusion of new authors and collaborations kept the diameter shrinking. In the third stage (2007-2013), the diameter was *stabilizing*, trying to reach an equilibrium.

4 Conclusions

We are interested in the arXiv GR-QC (General Relativity and Quantum Cosmology) network. However, existing data collections cannot satisfy our analysis needs. As a result, we created a data collection of the papers submitted from July 1992 to November 2013. One challenge of preparing a clean data collection is how to map author names to real authors. Even we tried our best to find a few rules to accomplish the task, we cannot find all of them with the limited amount of time. There is still room for improvement. We also applied various methods learned during the course to analyze and visualize the collaboration network. The scale-free network is also a small-world network with six degrees of separation. Most authors are contained in the giant component. The network also contains significant community structure. We visualized a few of them to gain some insight to the network. By studying how the graph evolved over year, we found the network was becoming denser, each author had more collaborators in average, and the diameter was shrinking.

References

[1] ArXiv General Relativity and Quantum Cosmology. `http://arxiv.org/archive/gr-qc`.

[9] *diameter* `http://igraph.sourceforge.net/doc/R/diameter.html`.

[2] Jure Leskovec. General Relativity and Quantum Cosmology collaboration. http://snap.stanford.edu/data/ca-GrQc.html.

[3] The Python Programming Language. http://www.python.org.

[4] M. E. J. Newman. The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences. PNAS, vol. 98, no. 2, 404-409, January 16, 2001. http://www.pnas.org/content/98/2/404.full.

[5] R Core Team. The R Project for Statistical Computing. http://www.python.org.

[6] Coefficient of Determination. http://en.wikipedia.org/wiki/Coefficient_of_determination.

[7] Gephi. https://gephi.org.

[8] Christian Lorenz Staudt. Analysis of Scientific Collaboration Networks: Social Factors, Evolution, and Topical Clustering. 2011. http://i11www.iti.uni-karlsruhe.de/_media/teaching/theses/da-staudt-11.pdf.

[9] B.Sathyaprakash, et al. Scientific Potential of Einstein Telescope. Gravitational Waves and Experimental Gravity, 2011. http://arxiv.org/abs/1108.1423.

[10] The Home Page of Dr. Lorenzo Iorio. http://digilander.libero.it/lorri/homepage_of_lorenzo_iorio.htm.

[11] Matt Visser's personalized homepage. http://homepages.mcs.vuw.ac.nz/~visser/.

[12] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-Law Distributions in Empirical Data. SIAM Review 51(4), pages 661-703, 2009. http://arxiv.org/pdf/0706.1062.pdf.

[13] Laurent Dubroca. R Code for Fitting a Power-Law Distribution. http://tuvalu.santafe.edu/~aaronc/powerlaws/plfit.r. The code is the same with the one used for the optional programming assignment 2, Social Network Analysis. https://spark-public.s3.amazonaws.com/sna/other/R/OPA2Template.R.

[14] Connected Component (graph theory). http://en.wikipedia.org/wiki/Connected_component_(graph_theory).

[15] Charu C. Aggarwal. Social Network Data Analytics (1st ed.). Springer Publishing Company, Incorporated, 2001.

[16] Small-world Network. http://en.wikipedia.org/wiki/Small-world_network.

[17] Six Degrees of Separation. http://en.wikipedia.org/wiki/Six_degrees_of_separation.

[18] Clustering Coefficient. http://en.wikipedia.org/wiki/Clustering_coefficient.

[19] Erdős - Rényi Model. http://en.wikipedia.org/wiki/Erd%C5%91s%E2%80%93R%C3%A9nyi_model.

[20] A Clauset, MEJ Newman, and C Moore. Finding community structure in very large networks. Physical Review E, 2004. http://www.arxiv.org/abs/cond-mat/0408187.

[21] Newman, M. E. J., Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E, 2001.

[22] Fruchterman Reingold Layout. http://wiki.gephi.org/index.php/Fruchterman-Reingold.

[23] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (KDD '05), 2005.

[24] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. ACM Trans. Knowl. Discov. Data 1, 1, Article 2, March 2007.