ATLANTIS PRESS

Atlantis Highlights in Computer Sciences, volume 2

Fourth Workshop on Computer Modelling in Decision Making (CMDM 2019)

Random Graph Models and Their Application to Twitter Network Analysis

Kirill Shaposnikov, Irina Sagaeva*, Alexey Grigoriev, Alexey Faizliev[0000−0001−6442−4361], Andrey Vlasov
Saratov State University, Saratov, Russian Federation
*sagaevaid@gmail.com

***Abstract*—In this paper, we conducted an experiment for comparison of the graphs generated by Erdős-Rényi, Barabási-Albert, Bollobás-Riordan, Buckley–Osthus, Chung-Lu models and a web graph constructed using real data. Twitter data have been employed to construct social network, and C++ has been used for network analysis as well as network visualization. It was shown that distribution of degrees and clustering coefficient for this network follows the power law. A machine learning approach is used for empirical evaluation of the Erdős-Rényi, Barabási-Albert, Bollobás-Riordan, Buckley–Osthus, Chung-Lu models in comparison to the Twitter graph.**

***Index Terms*—social network analysis, classification, degree distribution, social graph**

I. Introduction

One of the most important problems in the modern social sciences is the search for effective ways of summarizing and visualizing the data of social interactions of individuals. This provides useful information about the behavior of people within the communities to which they belong. Currently, the social interaction of individuals has become possible to study on the basis of data of their interactions in social networks such as Twitter, Facebook etc. The amount of data generated by social networks every day is huge. The analysis of such data becomes more complex as the number of interacting individuals and the amount of their interactions grow.

Twitter data can be easily transformed into network data. A social network is a set of individuals which have connections in pair to represent their relationship. Two individuals can be considered in a relationship if there has been friendship or common interest between them. In such type of network, an individual will be called as "node" or "vertex" and the connection will be an "edge". In the case of Twitter network, which is studied in this paper, the social network will be represented by directed unweighted graph in which links stand for retweets. If someone uses Twitter's Retweet icon then their Retweet or Retweet with comment will reference the tweet that they are sharing. Different types of social network analysis metrics can be used for finding e dge d ensity, degree distribution, maximum clique and maximum independent set in the network. This methodology allows us to visualize a data set, presenting its elements as vertices, and observe certain relationships between them. Studying the structure of a graph representing the data is important for understanding the intrinsic properties of social interaction.

Note that the last two decades have seen extensive research in the area of degree distribution analysis of complex networks arisen in sociology, physics, and biology. It has been shown that many networks have similar degree distributions [1]–[6]. It turned out that most real networks have degree distributions that are scale-free [1]. In other words, their degree distributions follow power law. Barabasi and Albert proposed the preferential attachment model in papers [7], [8] to fit the power law degree distribution of real web-graphs.

One of the most well-known social networks is Twitter which was created in 2006. Twitter which uses its minimalist concept of microblogging is an online application that aims at bringing together hundreds of millions of users around the world to express themselves. Each message (tweet) should have up to 140 characters and can be "followed" by other users without mandatory reciprocity, coupled to a very open application programming interface (API), which makes it an ideal tool for social behavior studies. In recent years, many researchers have studied the information dissipation which took place after such events as: an earthquake [9], political demonstrations [10], [11], teachings [12] or interactions on neutral corpus [13]. These studies usually visualize the networks of tweets as well as mentions and retweets. Twitter data is favored by researchers due to their availability and the possibility to inspect its contents. Social network analysis metrics allow to examine and measure the influence in political communication or scientometry [14], [15]. The paper [16] is aimed to find key factors to explain clustering within a network whose characteristics look similar to a small world.

In this paper we conducted an experiment for comparison of the graphs generated by Erdős-Rényi, Barabási-Albert, Bollobás-Riordan, Buckley–Osthus, Chung-Lu models and a web graph constructed using real data. Twitter data have been employed to construct social network, and C++ has been used for network analysis as well as network visualization. It was shown that distribution of degrees and clustering coefficient for this network follows the power law. A machine learning approach is used for empirical evaluation of the Erdős-Rényi, Barabási-Albert, Bollobás-Riordan, Buckley–Osthus, Chung-Lu models in comparison to the Twitter graph.

II. Model Comparison Methodology

One of the goals of the paper is to examine models for generating random graphs according to their suitability for fitting the Twitter user graph. In this paper we will use the approach proposed in works [17], [18]. The methodology consists of four steps:

1) We construct the Twitter user graph using real-world data. We also generate a set of graphs (subgraphs) from the real-world networks.
2) We select a set of models for generating random graphs. For each random graph model, we generate a collection of graphs. Thus, we obtain the collection of subgraphs taken from real-world Twitter network and the collection of subgraphs taken from the graphs constructed with the use of each generative random graph model.
3) We select a set of parameters which characterize networks. For all graphs in each collection we compute different subgraph parameters (e.g. its edge density, clustering coefficient, the subgraph diameter, different centrality measures, etc.). Thus, for each subgraph we obtain a feature representation vector. The collections of subgraphs are transformed into collections of feature representation vectors.
4) Then we perform a classification task to distinguish between these two sets of vectors representing any two generative models for random graphs using a machine learning algorithm. Then using the classifier we find the probabilities that subgraphs taken from real-world Twitter network belong to each of the two collections. The idea is that if the classifier can not distinguish two sets then the underlying random graph generating model is better in the sense of producing graphs with given characteristics.

In this paper we will use support vector machines (SVMs) with the Gaussian radial basis function (rbf) as our supervised machine learning classifier. The choice is due to the fact that the model has a good predictive performance [19] and is able to capture high order dependencies [20]. The classifier should return the probability that a random graph generated with use of some model is from the Twitter user graph.

Before solving the problem, it is necessary to process the collected graph, namely, to learn how to select from it coherent subgraphs of various sizes to compile a test data set. We also need to chose the set of graph characteristics (features). The set of characteristics should reflect the graphs to make their classification more accurate. In this paper we will employ the following characteristics that will be used to classify graphs:

- the number of vertices, n;
- the number of edges, m;
- the graph diameter, which is the maximum of the shortest distances between any two nodes;
- the effective diameter, which is defined as the minimum distance at which 90% of the nodes are reachable from each other;
- mean value, median, first and third quartiles, maximum and minimum values, standard deviation and variance, kurtosis coefficient, asymmetry coefficient of
  - clustering coefficient;
  - betweenness centrality;
  - closeness centrality;
  - Katz centrality;
  - PageRank coeffitient.

III. Twitter Graph

The web graph is a network in which each web page is considered as a node and each hyperlink is considered as an edge between two nodes. Perhaps, one of the first attempts to use a crawl of the web to examine some important characteristics of the web graph was performed in papers [8], [21]. This web graph consists of around 325,000 pages and more than 1.4 million edges. The main findings of this paper was that the degree distribution of the network follows the power law. This was quite surprising result at the time since power-law degree distribution can not be explained by classical random network models including Erdös–Rényi model [22], [23].

One of the first attempts to construct a random graph model fitting the power law distribution was the model proposed by Barabási and Albert. The model uses the idea of a so-called preferential attachment to naturally explain the growth of a graph. Barabási–Albert model captures such crucial properties of real networks as the logarithmic growth and the power law degree distribution. Moreover, Barabási and Albert were the first who associate the power law degree distribution (i.e. scale-free property of networks) with web graphs. A drawback of the model in describing real web graphs is that it is not realistic in capturing the true value of power law exponent. Moreover, it does not allow us to construct undirected graphs.

A crawler program was written in the Go language. It receives the user ID from which to start the crawl, as well as a number that determines how deep the crawl should be, because Twitter is too large and it is impossible to build a graph for all its users. To determine who reads tweets of a user, you can use the Twitter API, which has several advantages: reliability, good documentation, flexibility. However, it allows only a small number of requests in a period of time for a free account. Twitter API has a `friends/ids` HTTP endpoint to get the list of friends of the specified user with a limit of 15 requests per 15 minutes, i.e. one request per minute, which makes senseless parallelization of the code to increase performance. On the contrary, a special delay was provided in order to be within the allotted restrictions and not risk getting the application blocked. The downloaded friends list of the current user is stored in the NoSQL MongoDB database, since later it is convenient to get the graph sections without decoding the entire file, as would be the case with CSV format.

The parameters of the collected graph are presented in Table I.

Looking at these characteristics, we can conclude that the graph turned out to be fully connected, not dense and disassortative, i.e. the hubs are connected not directly, but through chains of neighbors. The exponent of the power law lies in the interval [2, 3], but this is not enough to say that the graph degree distribution follows the power law distribution. In order to check this in the `powerlaw` library there is a method `distribution_compare()` that allows you to compare distribution, which accepts string constants at the

TABLE I
TWITTER GRAPH CHARACTERISTICS.

Characteristics	Value
Vertices	13544931
Edges	34885363
Isolated vertices	0
Loops	0
Density	0.00000038
Clustering coefficient	0.084986
Minimum degree	1
Maximum degree	15476
Average degree	5.151058
Assortativity coefficient	-0.236831
# of connected components	2
The size of giant component	13544800 (100%)
Power law exponent	2.255

input corresponding to the distributions. The result of the method is the likelihood coefficient, if it is greater than 0, then the first distribution is preferable, otherwise the second one [24]. Table II presents the results obtained for the collected graph.

TABLE II
THE RESULTS OBTAINED BY FUNCTION DISTRIBUTION_COMPARE() WHEN APPLYING TO THE DISTRIBUTIONS 1 AND 2

Distribution 1	Distribution 2	Likelihood coefficient
Power low	Exponential	2216778.52
Power low	Truncated power law	0.00016
Power low	Lognormal	362.50

All the likelihood coefficients are greater than 0, therefore the graph degree distribution follows the power law distribution.

IV. MODELS

A. Erdős-Rényi Model

The most well-known model for generating random graphs was introduced by Erdős-Rényi. We chose its variant $G(n,p)$ based on connecting nodes randomly, the method was proposed by Edgar Gilbert. In an n-nodes graph, every node in it with probability p is connected to another node in the graph.

B. Barabási-Albert Model

The previous model does not exhibit power law. The Barabási-Albert model, however, generates scale-free networks with the use of preferential attachment. While starting as an m-nodes connected network, we add new nodes by a specific rule. Each new node is connected to a number of nodes by a certain rule, explained by the equation:

$$p_i = \frac{k_i}{\sum_j k_j}.$$

C. Bollobás–Riordan Model

The random graph constructed with the use of preferential attachment model proposed by Barabási and Albert [8], [21] exhibits the logarithmic growth that delivers the anticipated power law degree distribution. While the model is simple, it may poorly describe the real dynamics of the network. B. Bollobás and O. Riordan suggested an improved preferential attachment model in which exact formulas for obtaining the degree distribution, the diameter and the clustering coefficient were presented [25]. To better describe the model, the concept of linearized chord diagrams (LCDs) was introduced by B. Bollobás and O. Riordan [25].

Let n be the number of vertices and m be the parameter describing the edges to vertices number ratio, $n, m \in \mathrm{N}$. Bollobás–Riordan model [25] for generating a random graph G_m^n consists of two stages. At the first stage we should construct the graph G_1^n by induction. It is supposed that the graph G^{n-1} is already constructed. We add a new vertex n with an edge conducted by a certain rule to the G^{n-1} graph. We assume that we have the graph G_1^1 with one vertex and one edge loop, i.e. $G_1^1 = (\{1\}, \{(1,\, 1)\})$. Then in this graph we add a second vertex and draw an edge from the added vertex to the existing one with probability $\frac{\deg v}{2n-1}$, where $\deg v$ is the degree of i-th vertex, $i \in \{1, \ldots, n-1\}$, or build a loop in the added vertex with probability $\frac{1}{2n-1}$. The denominator is $2n-1$ for the reason that the sum of all probabilities must be equal to one. The constructed sequence of graphs is random.

In general, the probability distribution can be written as

$$P(i=s) = \begin{cases} \frac{d_{G_1^{n-1}}(s)}{2n-1},\ 1 \le s \le n-1; \\ \frac{1}{2n-1},\ s = n. \end{cases} \quad (1)$$

After the graph G_1^n has been obtained, we proceed to the second stage, where it is necessary to "collapse" its vertices. To do this, we will combine the vertices into groups of m vertices and each such group will be given its own notation v_i, where $i \in \{1, \ldots, n\}$, for example, the vertex $v_1 = \{1, \ldots, m\}$, and the vertex $v_2 = \{m+1, \ldots, 2m\}$. At each vertex v_i we construct as many loops as the edges were between the vertices of the corresponding group of the original graph. Similarly, we construct edges between vertices v in such a way that an edge is drawn in the case if there is an edge between the corresponding groups of the original graph. Thus, the graph G_1^{mn} is transformed to the graph G_m^n.

D. Buckley–Osthus Model

Bollobásh-Riordan model is aimed to model a real graph the degree distribution of which follows the power law distribution, i.e the number of vertices with degree d is well approximated by $d^{-\gamma}$, where $\gamma = 3$. However, the model sometimes does not produce values of interest that are close to the real data seen through empirical studies. For example, the model does not fit to the real web graph for which $\gamma = 2.1 \pm 0.1$. Thus, the model requires some improvements. One of the proposed solutions was called Buckley–Osthus model [26]. This model introduces into Bollobásh–Riordan

TABLE III
THE PAIRWISE MODEL COMPARISON

	Erdős–Rényi	Barabási–Albert	Bollobás–Riordan	Buckley–Osthus	Chung–Lu
Erdős-Rényi	-	0.36	0.30	0.48	0.29
Barabási-Albert	0.64	-	0.36	0.54	0.24
Bollobás-Riordan	0.69	0.64	-	0.73	0.41
Buckley–Osthus	0.52	0.46	0.27	-	0.26
Chung-Lu	0.71	0.76	0.59	0.74	-

model a positive coefficient a independent of the degree and which is called initial attractiveness of a vertex. In this case, the algorithm for adding new vertices and edges to the generated graph does not change, only the probabilities of holding an edge to a new vertex begin to be calculated taking into account the coefficient a using equation (2):

$$P(i=s)=\begin{cases}\frac{d_{H_{a,1}^{n-1}}(s)+a-1}{(a+1)n-1}, & 1\le s\le n-1;\\ \frac{a}{(a+1)n-1}, & s=n.\end{cases} \quad (2)$$

E. Chung–Lu Model

Chung–Lu model is designed to model sparse massive graphs which degree distributions satisfy a power law. It was first used to simulate telecommunication network precisely. The model has two parameters: the logarithm of the size of the graph α and log-log growth rate of the graph β. The vertices of graph based on Chung–Lu model satisfy equations [27]:

$$\log y=\alpha-\beta\log x$$

or

$$|\{v:degv=x\}|=y=\frac{e^{\alpha}}{x^{\beta}}.$$

V. EMPIRICAL RESULTS

A. Comparison

We will compare the real Twitter graph with the models described in Section IV, i.e.

- Erdős–Rényi model,
- Barabási–Albert model,
- Bollobás-Riordan model,
- Buckley–Osthus model,
- Chung–Lu model.

In the section we will find the model which shows the best results of pairwise comparison, using methodology described in II.

The main idea is to construct the characteristics vector for each graph, created using the chosen model. Then the vectors are passed to classifiers for training purposes. It is followed by its predictions with the real-graph characteristics vector input.

Model classification can be achieved using `sklearn` python library, which holds algorithms and methods for data analysis and machine learning [28]. In our case, we need SVM classifier which is included in the library. Characteristic vectors are passed into classifier for binary classification. It is known that the results are more precise in case of two classes for SVM to distinguish. Moreover, SVM reacts heavily on the scale of characteristics, hence we normalize data beforehand. We use `StandardScaler` helper class in `sklearn` library to normalize data [28].

After training process we pass characteristic vectors of real Twitter subgraphs. The output of classifier is a probability that the vector belongs to the class of a random graph model. The classifier results are shown in Table III.

VI. CONCLUSION

Thus, we conducted an experiment for comparison of the graphs generated by Erdős–Rényi, Barabási–Albert, Bollobás–Riordan, Buckley–Osthus, Chung–Lu models and the web graph constructed using Twitter data. The results presented in Table III show that the best fit of the real web graph corresponds to the the graphs generated with use of the Chung–Lu model.

ACKNOWLEDGMENTS

This work was supported by Russian Foundation for Basic Research, project 18-37-00060.

REFERENCES

[1] R. Albert and A.-L. Barabasi, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, pp. 47–97, 2002.
[2] S. N. Dorogovtsev and J. F. F. Mendes, "Evolution of networks," *Adv. Phys*, vol. 51, p. 1079, 2002.
[3] M. E. J. Newman, "The structure and function of complex networks," *Siam Review*, vol. 45, pp. 167–256, 2003.
[4] R. Albert, "Scale-free networks in cell biology," *Journal of Cell Science*, vol. 118, pp. 4947–4957, 2005.
[5] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang, "Complex networks: Structure and dynamics," *Physics Reports*, vol. 424, pp. 175–308, 2006.
[6] C. Lofdahl, E. Stickgold, B. Skarin, and I. Stewart, "Extending generative models of large scale networks," *Procedia Manufacturing*, vol. 3, no. Supplement C, pp. 3868–3875, 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015.
[7] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999. [Online]. Available: https://science.sciencemag.org/content/286/5439/509
[8] A.-L. Barabási, R. Albert, and H. Jeong, "Mean-field theory for scale-free random networks," *Physica A: Statistical Mechanics and its Applications*, vol. 272, no. 1, pp. 173 – 187, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378437199002915
[9] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 851–860. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772777

[10] M. Beguerisse-Díaz, G. G. no Hernández, B. Vangelov, S. N. Yaliraki, and M. Barahona, "Interest communities and flow roles in directed networks: the twitter network of the uk riots," *Journal of The Royal Society Interface*, vol. 11, no. 101, p. 20140940, 2014. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2014.0940

[11] A. A. Casilli and P. Tubaro, "Social media censorship in times of political unrest - a social simulation experiment with the uk riots," *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, vol. 115, no. 1, pp. 5–20, 2012. [Online]. Available: https://doi.org/10.1177/0759106312445697

[12] C. Ullrich, K. Borau, and K. Stepanyan, "Who students interact with? a social network analysis perspective on the use of twitter in language learning," in *Sustaining TEL: From Innovation to Learning and Practice*, M. Wolpers, P. A. Kirschner, M. Scheffel, S. Lindstaedt, and V. Dimitrova, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 432–437.

[13] D. Darmon, E. Omodei, and J. Garland, "Followers are not enough: A multifaceted approach to community detection in online social networks," *PLOS ONE*, vol. 10, no. 8, pp. 1–20, 08 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0134860

[14] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network," in *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, ser. SOCIALCOM '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 177–184. [Online]. Available: https://doi.org/10.1109/SocialCom.2010.33

[15] S. Haustein, I. Peters, C. R. Sugimoto, M. Thelwall, and V. Larivière, "Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature," *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 656–669, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23101

[16] M. Grandjean, "A social network analysis of twitter: Mapping the digital humanities community," *Cogent Arts & Humanities*, vol. 3, no. 1, 2016. [Online]. Available: http://doi.org/10.1080/23311983.2016.1171458

[17] N. Attar and S. Aliakbary, "Classification of complex networks based on similarity of topological network features," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 9, p. 091102, 2017. [Online]. Available: https://doi.org/10.1063/1.4997921

[18] T. Bläsius, T. Friedrich, M. Katzmann, A. Krohmer, and J. Striebel, "Towards a systematic evaluation of generative network models," in *Algorithms and Models for the Web Graph*, A. Bonato, P. Prałat, and A. Raigorodskii, Eds. Cham: Springer International Publishing, 2018, pp. 99–114.

[19] M. F. Delgado, E. Cernadas, S. Barro, and D. G. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *J. Mach. Learn. Res.*, vol. 15, pp. 3133–3181, 2014.

[20] K. P. Bennett and C. Campbell, "Support vector machines: Hype or hallelujah?" *SIGKDD Explorations*, vol. 2, pp. 1–13, 2000.

[21] R. Albert, H. Jeong, and A.-L. Barabási, "Internet: Diameter of the world-wide web," *Nature*, vol. 401, no. 6749, pp. 130–131, Sep. 1999.

[22] P. Erdös and A. Rényi, "On random graphs i," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

[23] ——, "On the evolution of random graphs," *Publ. Math. Inst. Hungary. Acad. Sci.*, vol. 5, pp. 17–61, 1960.

[24] J. Alstott, E. Bullmore, and D. Plenz, "powerlaw: A python package for analysis of heavy-tailed distributions," *PLOS ONE*, vol. 9, no. 1, pp. 1–11, 01 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0085777

[25] B. Bollobás and O. Riordan, "The diameter of a scale-free random graph," *Combinatorica*, vol. 24, pp. 5–34, 2004.

[26] P. G. Buckley and D. Osthus, "Popularity based random graph models leading to a scale-free degree sequence," *Discrete Mathematics*, vol. 282, no. 1, pp. 53 – 68, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0012365X03006940

[27] B. Bollobás and O. Riordan, "The diameter of a scale-free random graph," *Experimental Math*, vol. 10, pp. 53–66, 2001.

[28] A. Gron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1st ed. O'Reilly Media, Inc., 2017.