

Online Social Networks and Media

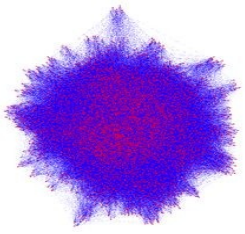
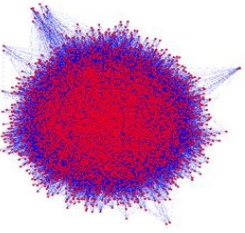
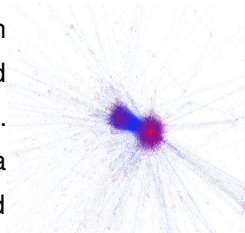
Assignment 1

Full Name(s) – Student ID(s): Terizi Chrysoula – 430, Triantali Dimitra – 431

Date: 13 – 11 – 2019

Question 1

The given Wiki-Vote network is characterized by ~7k wikipedia users and ~103k directed connections between them. Implementations that provided by SNAP and NetworkX libraries were used for the generation of networks. Specifically,

- **Erdos-Renyi model:** NetrowkX library provides the function `erdos_renyi_graph(n, p, seed=None, directed=False)`, which chooses each of the possible edges with probability p .
Tuning parameters: $n=7115$, $p=0.0041$. The model generates an undirected random graph which consists of ~7k nodes and ~103k edges. Based on Figure 1(a), one of the best value for the parameter p which achieves the minimum deviation of the Wiki-Vote number of edges is the selected one as mentioned above.
- **Preferential attachment model:** NetrowkX library provides the function `barabasi_albert_graph(n, m, seed=None)`, which the nodes are grown by attaching new nodes each with m edges that are preferentially attached to existing nodes with high degree². In real networks new nodes tend to link to the more connected nodes³.
Tuning parameters: $n=7115$, $m=15$. The model generates a network which consists of ~7k nodes and ~106k undirected connections. As can see in Figure 1(b), $m=14$ and $m=15$ are the dominant values to satisfy the requirement of creating a graph with total number of edges equals to this one of Wiki-Vote graph.
- **Forest Fire model:** SNAP library gives the opportunity to its users to use the function `GenForestFire(n, FwdProb, BckProb)`, which generates a random Forest Fire, directed graph with given forward and backward probabilities of an edge⁴.
Tuning parameters: $n=7115$, $FwdProb=0.73$ and $BckProb=0.08$. The model generates a graph of ~7k users and ~92k edges. Figure 1(c) shows that a linear pattern is observed between forward and backward probabilities in case where the subtraction of the number of edges of Wiki-Vote graph from the corresponding number of edges of generated Forest Fire graph. There are ~29 different pairs of (forward probability, backward probability) which the absolute distance between the number of edges less than 10000. As a consequence the selection of the parameters' values was based on this criterion.

1 https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.generators.random_graphs.erdos_renyi_graph.html

2 https://networkx.github.io/documentation/latest/reference/generated/networkx.generators.random_graphs.barabasi_albert_graph.html

3 <https://www.sciencedirect.com/topics/computer-science/preferential-attachment>

4 <https://snap.stanford.edu/snappy/doc/reference/GenForestFire.html>

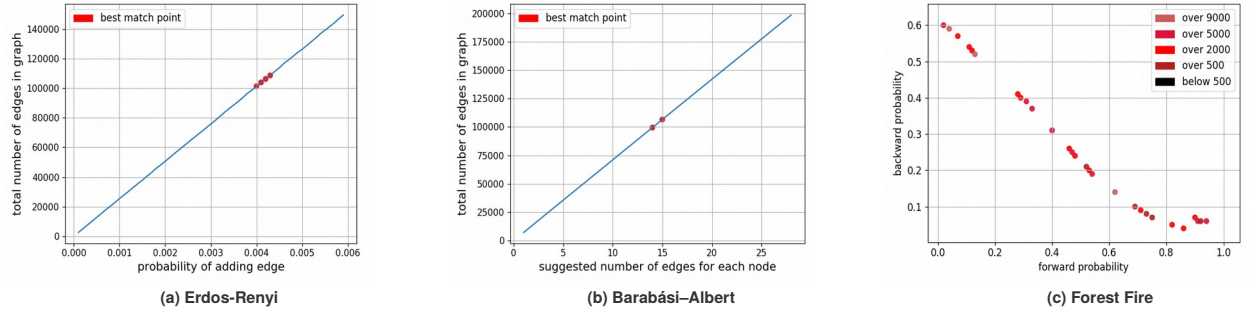


Figure 1: (a) Plot for each potential probability value which adds new edges, the average of 5 times of the total number of edges for Erdos-Renyi graph, (b) Plot for various values of the suggested number of edges for new nodes, the average of 5 times of the total number of edges for Barabási-Albert graph. The red points are the best points where the absolute subtraction value of the number of edges of Wiki-Vote graph from the respectively number of (a) and (b) graphs is less than 5000. (c) Subtraction of the number of edges of Wiki-Vote graph from the number of edges of Forest Fire graph for each combination of forward and backward probabilities.

Degree distribution plots

In real networks, the network's degree follows a power-law distribution. In a simple log-log representation of this distribution, noise is produced for the high degree values. As shown in Figures 2(a), (c) and (d), the simple degree distribution plot for real networks has this feature. The Erdos-Renyi graph is an absolutely random graph and as observed in Figure 2(b), it does not have the same behavior.

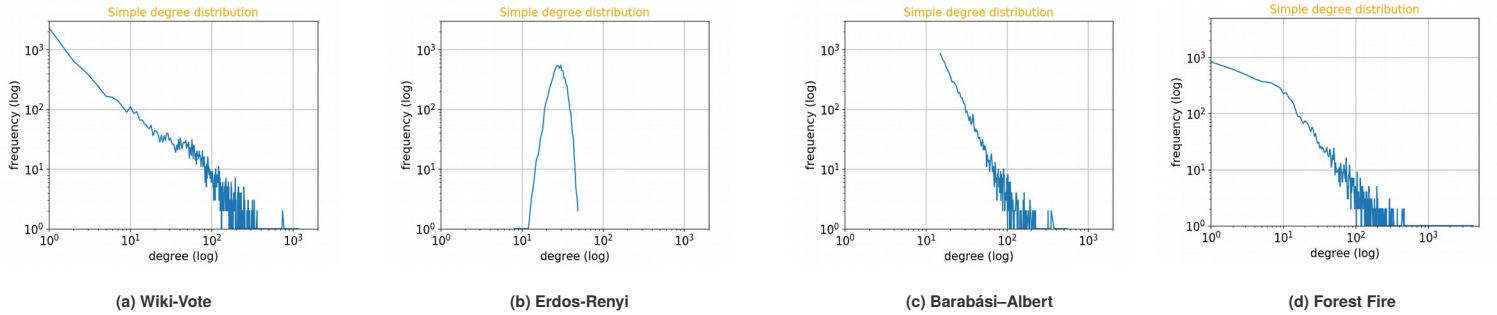


Figure 2: Simple (log-log) degree distribution plots for the Wiki-Vote network and the 3 generated networks: (a) Wiki-Vote, (b) Erdos-Renyi, (c) Barabási-Albert, and (d) Forest Fire.

The exponential binning representation of degree distribution creates bins that grow exponentially in size and significantly reduces the noise at the tail but not entirely. Figure 3 shows this elimination of noise except for the random Erdos-Renyi graph.

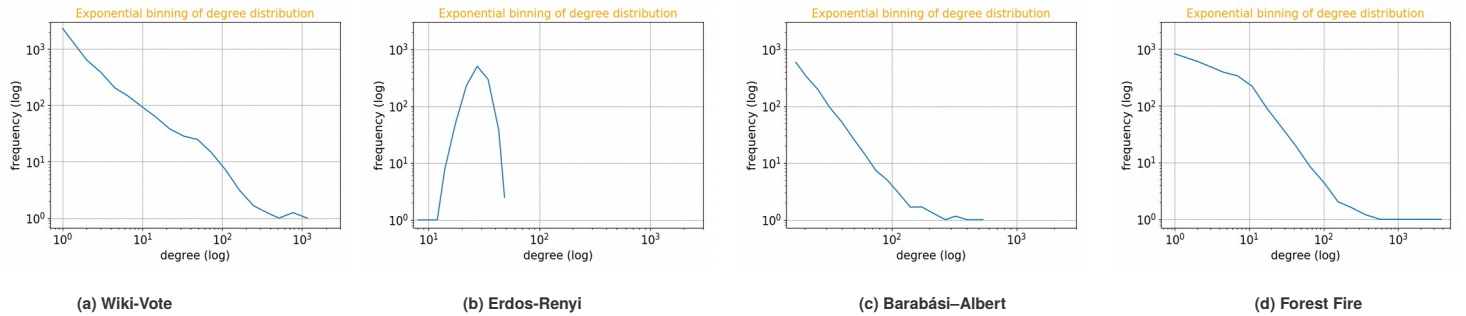


Figure 3: Logarithmic binning (log-log) degree distribution plots for the Wiki-Vote network and the 3 generated networks: (a) Wiki-Vote, (b) Erdos-Renyi, (c) Barabási-Albert, and (d) Forest Fire using 20 bins.

Observing the cumulative degree distribution for the real Wiki-Vote network (see Figure 4(a)) and the generated graphs (see Figures 4(c)-(d)), the probability of a node having a low degree is high over 0.80. For a random Erdos-Renyi graph (see Figure 4(b)), the probability of a node having low or high degree is uniform.

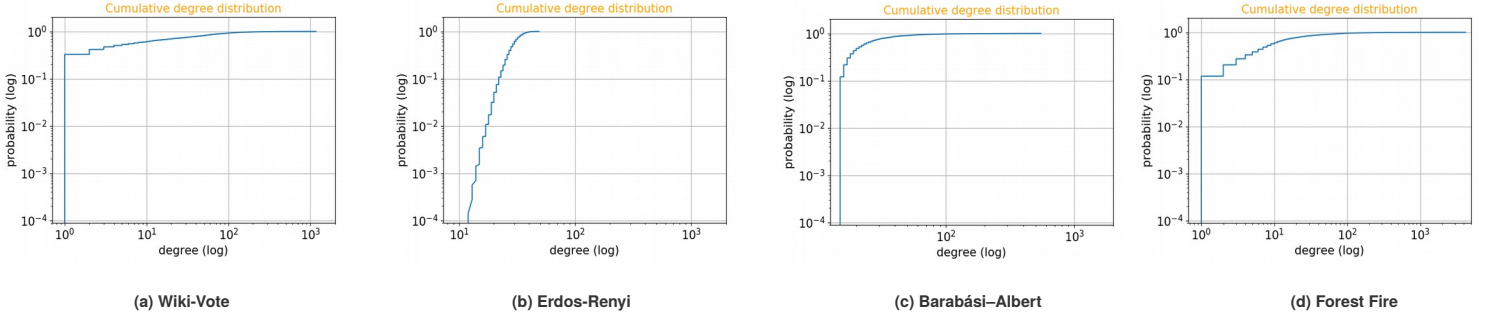


Figure 4: Cumulative (log-log) degree distribution plots for the Wiki-Vote network and the 3 generated networks: (a) Wiki-Vote, (b) Erdos-Renyi, (c) Barabási-Albert, and (d) Forest Fire.

About Zipf representation of degree distribution, if there is a power-law, the Zipf line is like a straight line. Figures 5(a), (c) and (d) of the real graphs tend to look like straight lines.

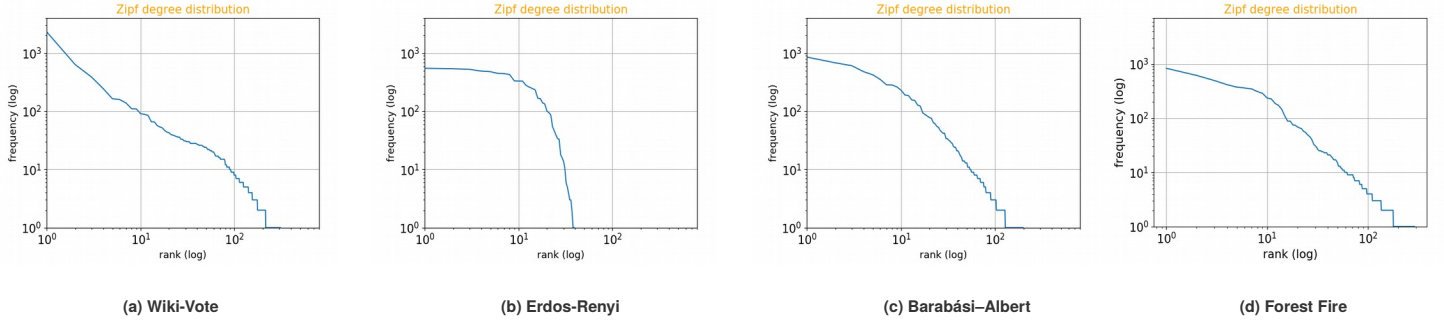


Figure 5: Zipf (log-log) degree distribution plots for the Wiki-Vote network and the 3 generated networks: (a) Wiki-Vote, (b) Erdos-Renyi, (c) Barabási-Albert, and (d) Forest Fire.

Effective diameter plots

The SNAP library provides the function `GetBfsEffDiam(Graph, NTestNodes, IsDir=False)`, which takes as input the SNAP graph and returns the (approximation of the) effective diameter (90-th percentile of the distribution of shortest path lengths) of a graph (by performing BFS from `NTestNodes` random starting nodes)⁵. Figure 6 shows that the effective diameter for an undirected graph (Erdos-Renyi and Barabási-Albert) is independent of the number of starting nodes and it is below 3 otherwise taking the strong connected component of directed (Wiki-Vote and Forest Fire) graphs, the number of starting nodes it matters. If the number of starting nodes exceeded 200, the effective diameter varies from 2.5 to 3.0 and from 3.5 to 4.0 respectively.

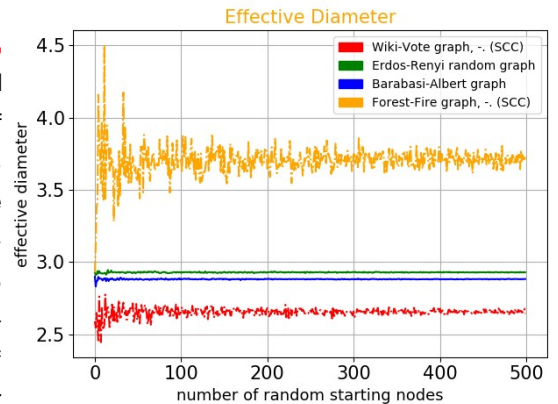


Figure 6: Approximation of the effective diameter for the Wiki-Vote graph and the 3 generated graphs.

⁵ <https://snap.stanford.edu/snappy/doc/reference/GetBfsEffDiam.html>

Clustering coefficient

Based on Gephi statistic tool “**Avg. Clustering Coefficient**”, the clustering coefficients are the followings in increasing order,

Order	Networks	Clustering coefficient
1	Erdos-Renyi	0.004
2	Barabási-Albert	0.018
3	Wiki-Vote	0.081
4	Forest Fire	0.257

PageRank plots

Good authorities should be pointed by good authorities. The authority value of each node is the sum of the authority fractions it collects from its neighbors. NetworkX library provides **pagerank(G, alpha=0.85, personalization=None, max_iter=100, tol=1e-06, nstart=None, weight='weight', dangling=None)** function which computes a ranking of the nodes in the graph G based on the structure of the incoming links⁶. Focusing on Figures 7(e)-(f), the 95 percentile of nodes have low pagerank values for both of the directed graphs. In 2007, Litvak, Scheinhardt & Volkovich⁷, prove that the pagerank values follow

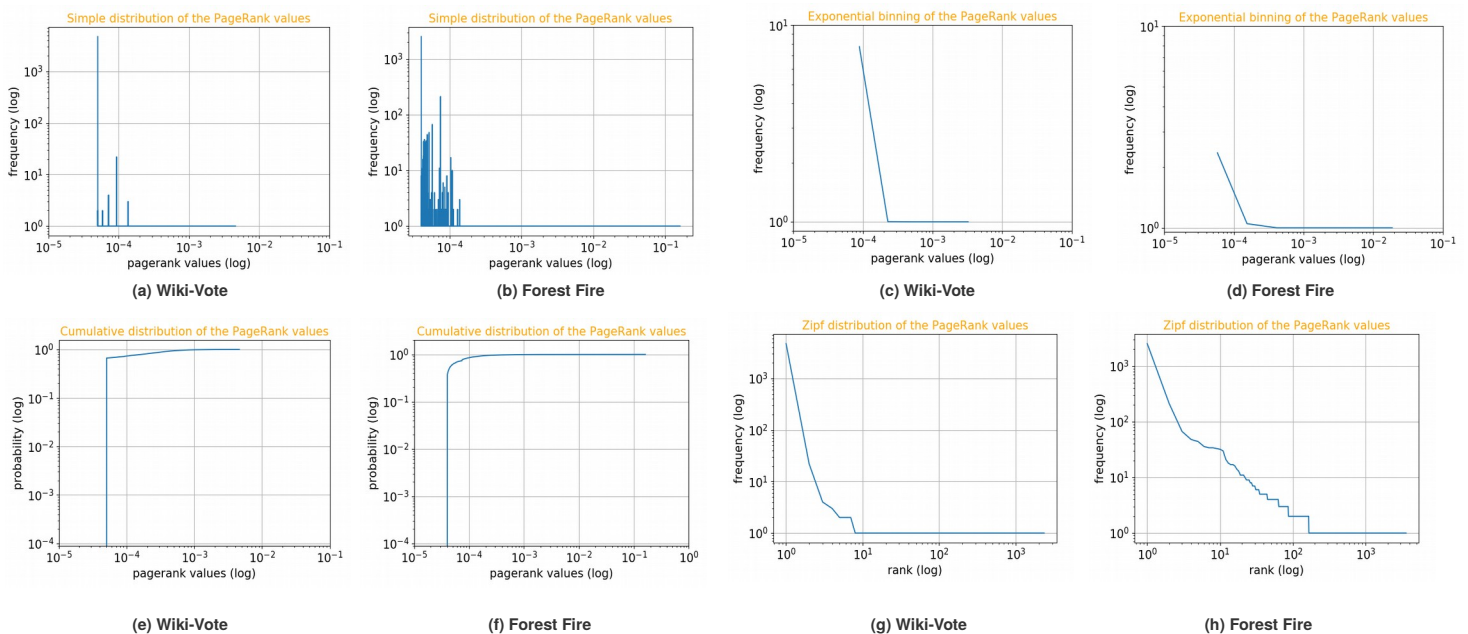


Figure 7: Simple (log-log) distribution plots of the pagerank values of the (a) Wiki-Vote and (b) Forest Fire networks. Logarithmic binning (log-log) distribution plots of the pagerank values of the (c) Wiki-Vote and (d) Forest Fire networks using 20 bins. Cumulative (log-log) distribution plots of the pagerank values of the (e) Wiki-Vote and (f) Forest Fire networks. Zipf (log-log) distribution plots of the pagerank values of the (g) Wiki-Vote and (h) Forest Fire networks.

similar power laws. Forest Fire graph shows up noise at the tail (see Figure 7(b)) which corresponds to the straight line in Figure 7(h).

⁶ https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.link_analysis.pagerank_alg.pagerank.html

⁷ N. Litvak, W. R. W. Scheinhardt & Y. Volkovich (2007) In-Degree and PageRank: Why Do They Follow Similar Power Laws?, Internet Mathematics, 4:2-3, 175-198, DOI: 10.1080/15427951.2007.10129293

Question 2

Given a network G which consists of $N \in \mathbb{N}$ number of nodes. Let p_v be the $N \times 1$ pagerank vector and $v = \left[\frac{1}{N} \quad \frac{1}{N} \quad \frac{1}{N} \dots \frac{1}{N} \quad \frac{1}{N} \right]^T$ the $N \times 1$ uniform jump vector. p_v is defined as follows:

- $p^0 = v$
- $p^1 = a p^0 P' + (1-a)v$
- \dots
- $p^\infty = (1-a)v(I - aP')^{-1}$

, where P is the $N \times N$ transition matrix which equals $P' = P + \vec{d} v^T$. Let \vec{d} be a $N \times 1$ vector where $d_i = 1$ if i is a sink node and $d_i = 0$ otherwise.

Let p_i be the $N \times 1$ personalized pagerank vector $\forall i \in [1, N]$, which is defined as follows:

- $p^0 = r_i$
- $p^1 = a p^0 P' + (1-a)r_i$
- \dots
- $p^\infty = (1-a)r_i(I - aP')^{-1}$

, where r_i is the $N \times 1$ jump vector for node i which has all probability on the node i . This means that $r_i^T = [0 \dots 0 \dots 1 \dots 0]^T$ where the i -th position is equals to 1.

$$\begin{aligned}
 \text{Hence, } \sum_{i=1}^N v(i) p_i &= \sum_{i=1}^N \frac{1}{N} (1-a) r_i (I - aP')^{-1} \\
 &= \frac{1}{N} (1-a) (I - aP')^{-1} \sum_{i=1}^N r_i \\
 &= (1-a) (I - aP')^{-1} \left(\frac{1}{N} [1 \ 1 \dots 1 \ 1]^T \right) \\
 &= (1-a) (I - aP')^{-1} \left[\frac{1}{N} \quad \frac{1}{N} \quad \dots \quad \frac{1}{N} \quad \frac{1}{N} \right]^T \\
 &= (1-a) (I - aP')^{-1} v \\
 &= p_v
 \end{aligned}$$