Small-World Properties of Facebook Group Networks

Article *in* Complex Systems · September 2014

DOI: 10.25088/ComplexSystems.23.3.197

CITATIONS
14

READS
526

2 authors, including:

Jason Wohlgemuth
Wohlgemuth Tech Foundation
1 PUBLICATION **14** CITATIONS

Small-World Properties of Facebook Group Networks

Jason Wohlgemuth
Mihaela Teodora Matache*
University of Nebraska at Omaha, Mathematics
Durham Science Center 203, Omaha, NE 68182, USA
**Corresponding author, dmatache@unomaha.edu.*

Small-world networks permeate modern society. In this paper we present a methodology for creating and analyzing a practically limitless number of networks exhibiting small-world network properties. More precisely, we analyze networks whose nodes are Facebook groups sharing a common word in the group name and whose links are mutual members in any two groups. By analyzing several numerical characteristics of single networks and network aggregations, we investigate how the small-world properties scale with a coarsening of the network. We show that Facebook group networks have small average path lengths and large clustering coefficients that do not vanish with increased network size, thus exhibiting small-world features. The degree distributions cannot be characterized completely by a power law, and the clustering coefficients are significantly larger than what would be expected for random networks, while the average shortest paths have consistently small values characteristic of random graphs. At the same time, the average connectivity increases as a power of the network size, while the average clustering coefficients and average path lengths do not exhibit a clear scaling with the size of the network. Our results are somewhat similar to what has been found in previous studies of the networks of individual Facebook users.

1. Introduction

In the past few years there has been a great interest in studying the basic topology of a variety of networks such as the World Wide Web [1], signal transduction networks [2], subway systems [3], railway networks [4], and more recently, Facebook [5, 6]. These efforts have seen the emergence of a very specific type of network, the small-world network, in which most nodes are not neighbors of one another but can be reached in a small number of steps. The concept of a small world made its way into the realm of academia via the work of the social psychologist Stanley Milgram [7, 8]. Physicists entered the fray with the development of the aptly named Watts–Strogatz network

https://doi.org/10.25088/ComplexSystems.23.3.197

model [9], which replaced the previously uncontested random graph model of Erdős and Rényi [10]. In the Erdős–Rényi model, the links between nodes are generated independently with the same probability. The Watts–Strogatz model starts with a ring lattice, and each link is rewired with a given probability p. If $p = 1$, the Erdős–Rényi model is obtained, while $p = 0$ yields the ring lattice. Watts and Strogatz considered two numerical characteristics associated with the network: the characteristic path length and the clustering coefficient [9]. They showed that for very small values of p and for large enough values of p, the two numerical characteristics tend to have similar magnitudes. However, for intermediate values of p, the two numerical characteristics tend to be at opposite ends of their range, [0, 1], namely small characteristic path length and large clustering coefficient. That behavior corresponds to small-world networks. More recently, Serra et al. [11] introduced the equal number of links algorithm to generate small-world networks starting from a regular lattice, by randomly rewiring some connections. That study aimed at analyzing the dynamics of interacting oscillators or automata. It was found that key dynamical properties (i.e., number of attractors, size of basins of attraction) are modified by rewiring; for example, there is a decrease in the number of attractors that are reached. On the other hand, Aguirre et al. [12] described an algorithm for generating a small-world graph with a higher number of biconnected components than lattices have, which is useful for modeling hierarchical multi-agent networks or the internet. It was shown that these kinds of networks present a slower descent in their characteristic path length; however, no significant difference was observed in the clustering coefficient behavior in comparison to a ring-lattice approach.

The applications of these models are all around us: the neural networks in our brains, the ecosystems of rain forests, the future of the stock market, the dynamics of epidemics, and the internet, to name a few. For example, in a study by Barabási, Albert, and Hawoong [1], it was found that the World Wide Web is in fact a scale-free network; that is, the connectivity has a power-law distribution with a heavy tail. Later studies would find that several small-world networks have scale-free topologies. In [13] Albert and Barabási provide a comprehensive treatment of networks and include the parameters of many small-world networks of a diverse variety. Small-world networks have become quite mainstream with books such as Watts' *Six Degrees* [14], and Barabási's *Bursts* [15] and *Linked* [16]. Latora and Marchiori [3] made the leap from theory to experiment by considering a complex network where the nodes are train stations and the edges are stations connected by track. In their own terminology, they found the Boston subway to be locally and globally efficient, tantamount to be-

https://doi.org/10.25088/ComplexSystems.23.3.197

ing a small-world network. The Indian railway network was also found to exhibit small-world properties by Sen et al. [4].

This paper is concerned with characterizing real-world examples of the small-world network phenomenon, starting with features such as clustering, degree distribution, or average path length for one of the most expansive networks: Facebook. Previous analysis of Facebook has focused on individuals [5, 6]. The work in [5] confirms "six degrees of separation" to be present in the Facebook graph, along with high local clustering. That is, Facebook, at the level of individual users, seems to be a small-world network, and the degrees of separation would later be reduced to four [6]. We supplement the previous research by focusing on groups rather than individuals. There is much to be learned about networks by studying expansive social networks such as Facebook, but small-world properties have been seen and analyzed in networks with as few as 43 nodes [17]. We are interested in understanding the impact of aggregation of individuals into groups on the main statistics of the network. It has been noted that the nature of small-world graphs makes it difficult for many coarsening approaches to retain the relevant properties of the original graph [18]. In order to be able to perform a comparison with the previous results in the literature, we use similar topology statistics [13, 19] and information on connectivity or degree distribution, clustering coefficients, average shortest paths, and network density. We use the average shortest paths and clustering coefficients to characterize small-world networks, as is done in [9, 17]. In this regard, we create several networks composed of Facebook groups with a common keyword in their titles and compare the parameters of each to find common characteristics and observe how they change with other parameters such as, but not limited to, the number of nodes, graph density, and degree distribution. We compare our results to the corresponding ones for random networks or other Facebook studies. We show that an aggregation of Facebook users into groups and of groups into further smaller categories of groups does not change the basic small-world features. Thus Facebook exhibits a scaling invariance of properties. At the same time, we note some differences between networks representing different interests, such as politics versus sports.

Basically, in this paper we generate a coarsening of the Facebook network by classes of personal interests of Facebook users based on a number of keywords for each area of interest under consideration. The links are generated by individuals common to different keyword groups. This type of coarsening is shown to preserve small-world features that have been noted in [5, 6] at the Facebook user level. Similarly to [5], we show that a strict power law may not be the best fit for the degree distribution of the Facebook network. Thus, this prop-

https://doi.org/10.25088/ComplexSystems.23.3.197

erty is also preserved by the coarsening of the network. We show that the average connectivity increases as a power of the network size with approximation, while the average clustering coefficients and average path lengths do not exhibit a clear scaling with the number of nodes of the network.

There are two complementary elements to this paper: the description of the social network under consideration and the mathematical network analysis. Although the first element will be described in sufficient detail, the main focus is the mathematical analysis.

The organization of this paper is as follows. Section 2 lays the foundation for the mathematics needed for analysis. Section 3 describes in detail how we create our networks from the data pulled from Facebook. We present the results of our analysis, including visualizations and statistical approaches, in Section 4 and end with conclusions and ideas for future work in Section 5.

2. Mathematical Background

In this section we provide a brief overview of the mathematical tools needed to analyze the networks in this paper. We also review the numerical characteristics of random and small-world networks.

2.1 The Facebook Group Network and Significant Numerical Measures

Let us denote by $G = \{x_1, x_2, \ldots, x_N\}$ a network with N nodes. Each node x_i is assumed to be linked to $k_i \in \{0, 1, \ldots N-1\}$ other nodes in the network, called its inputs or neighbors. The parameter k_i is called the connectivity or degree of node x_i. If $k_i = 0$, then the node is isolated. Here we deal with undirected networks; that is, if node x_j is an input to node x_i, then x_i is an input to x_j. G can be viewed as a graph with vertices $x_1, x_2, \ldots, x_N$ and edges (x_i, x_j), $i, j = 1, 2, \ldots N$.

The actual Facebook network considered in this study is described as follows.

Definition 1. Let F_w be the set of (public) Facebook groups with the word w in the group's title. These groups represent the nodes of the network, and consequently each node x_i is basically the set of people belonging to that group. Then we can define the set of links/edges between the nodes of this network as

$$L_w = \left\{(x_i, x_j) \in F_w \times F_w \,\middle|\, x_i \cap x_j \neq \emptyset, i, j = 1, 2, \ldots, N, i \neq j\right\}.$$

Furthermore, two nodes x_i and x_j are said to be adjacent nodes if

https://doi.org/10.25088/ComplexSystems.23.3.197

$(x_i, x_j) \in L_w$. Thus, we construct G_w, a network with nodes F_w and edges L_w, as defined.

For example, if $w =$ food, then we consider all the groups that have the word food in their title. Two possible group titles could be "Food and wine" and "All about food." These would be two nodes in the network. If they share common members, then there is an edge between them. The main reason for constructing such networks is to be able to identify possible commonalities as well as differences between the structure and properties of networks that are all social, but represent different types of personalities and interests. We would expect to see some impact of the type of common interest of the groups, represented by the common word. For instance, we find that the $w =$ bieber (Justin Bieber, Canadian singer-songwriter, musician, producer, and actor, born 1994, http://en.wikipedia.org/wiki/Justin_Bieber) network is a lot more connected and clustered than, for example, the $w =$ math network, not to mention much, much larger. Often some differences are intuitive: the graph F_{bieber} is large and connected, while the graph F_{biology} is almost nonexistent. This is most likely due to the different populations represented by the two groups, as well as the types of personalities and personal interests of the individuals belonging to these groups: some are interested in being in a music group related to their personal preference for music, others in a biology group perhaps related to their professional life. The choice of keyword affects the return drastically, and the resultant graph is subject to various factors, both intuitive and otherwise. It is our goal to analyze the impact of these common interests on the numerical characteristics of Facebook group networks in order to decide if the networks possess small-world characteristics and to identify differences or similarities between groups with possibly unrelated interests, representing different segments of the population and different personalities. To this end, the most common numerical characteristics to be analyzed are the degree distribution, average clustering coefficient, and associated path lengths of the network [4–6, 9, 13]. We now recall their definitions.

Definition 2. If $k_1, k_2, \ldots, k_N$ are the connectivity values of the nodes $x_1, x_2, \ldots x_N$, respectively, the connectivity distribution is given by the probability distribution function $f(x) = P(k_i = x)$ for any i, where $x \in \{1, 2, \ldots N - 1\}$.

Definition 3. The clustering coefficient of a node x_i, denoted C_i, is a measure of transitivity. That is, it measures how connected the inputs of a node are. More precisely, if node x_i has k_i inputs, then there exist

https://doi.org/10.25088/ComplexSystems.23.3.197

at most $1/2\, k_i(k_i - 1)$ links between these k_i inputs. C_i is defined as the fraction of the number of the links that actually exist in the network with respect to the total number of possible links. Then the average clustering coefficient for a network with N nodes is

$$\overline{C} = \frac{1}{N} \sum_{i=1}^{N} C_i. \tag{1}$$

If $\hat{k} = \{x \in F_w \mid \text{degree}\,(x) = k\}$, where $k \in \{1, 2, \ldots, N-1\}$, then the average clustering coefficient of nodes with degree k is

$$\overline{C}(k) = \frac{1}{|\hat{k}|} \sum_{x \in \hat{k}} C_x, \;\; k = 1, 2, \ldots, N-1, \tag{2}$$

where $|\hat{k}|$ denotes the cardinality of set $\hat{k}$.

Definition 4. Given two nodes x_i and x_j, the shortest path connecting them, l_{ij}, is given by the minimal number of links that lead from x_i to x_j (or vice versa since we are dealing with an undirected network, so that $l_{ij} = l_{ji}$). The average path length is

$$\langle l \rangle = \frac{1}{N(N-1)} \sum_{i \neq j} l_{ij}. \tag{3}$$

By construction, we eliminated nodes with $k = 0$ (isolated nodes), which means each network is comprised of one or more subgraphs/subnetworks where every node in each subgraph has $k \geq 1$. In that case, the average path length of the network is determined as the mean of the average path lengths of each subgraph.

A small-world network is characterized by the following properties:

1. The average shortest path length scales with $\ln N$ [13].
2. The network exhibits clustering higher than random networks [13, 20].

Based on previous studies [9, 20, 21], these two requirements provide a good metric for assessing real-world networks that exist between complete order and randomness. An ordered network is basically a ring lattice in which the nodes are placed on a circle, and each node is connected to its k nearest neighbors. With this idea in mind, we provide a brief review of the models for random and small-world networks.

https://doi.org/10.25088/ComplexSystems.23.3.197

2.2 Random Graph Models

A random network is a network in which some specific parameters, such as the number of edges or the average connectivity, take fixed values, but the network is random in other respects. There are several models for generating random graphs, the most common one being to maintain the number of nodes and edges constant while randomly assigning the edges [22]. Arguably the most widely studied type of model for the construction of a random network is $G(N, p)$, where N is the number of nodes of the network, and p is the (fixed) probability of constructing an edge between any two nodes. These networks have become known as Erdős–Rényi networks [10], due to the eminent works of the namesakes. The connectivity for $G(N, p)$ networks follows a binomial distribution,

$$f_{\text{rand}}(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}. \tag{4}$$

The average clustering coefficient is

$$\overline{C}_{\text{rand}} = \frac{\langle k \rangle}{N}, \tag{5}$$

while the average shortest path length scales as follows [13]

$$l_{\text{rand}} \sim \frac{\ln N}{\ln \langle k \rangle}. \tag{6}$$

Therefore, for a $G(N, p)$ network, as $N \to \infty$, we see that $\overline{C}_{\text{rand}}$ vanishes and l_{rand} scales as $\ln N$. Thus, property 1 of small-world networks is fulfilled. However, despite the fact that there is a relatively small path length between any two nodes, the random graphs lack the inherent nontrivial clustering of a small-world network described by property 2.

2.3 Small-World Network Models

In an effort to capture the transitivity of actual small-world networks and retain average shortest paths characteristic of random graphs that scale with $\ln N$, the Watts–Strogatz model was developed [9]. More precisely, starting with a ring lattice, each edge is rewired with probability p, excluding self-inputs and duplicate edges. When there is no chance of an edge's being rewired $p = 0$, we obtain an ordered/regular network. It exhibits high clustering, fulfilling property 2, but has long path lengths, so it is not a small-world network. At $p = 1$, a Watts–Strogatz network is exactly a random network and thus has

https://doi.org/10.25088/ComplexSystems.23.3.197

short path lengths to fulfill property 1 but little to no transitivity, so it is not a small-world network. For a significant range of values of $0 < p < 1$, the networks created with the Watts–Strogatz model can fulfill both properties 1 and 2, so they are small-world networks.

The degree distribution for the Watts–Strogatz small-world model is Poisson with parameter $k\,p$ and the clustering coefficient is

$$C = \frac{3\,(k-2)}{4\,(k-1) + 8\,k\,p + 4\,k\,p^2}, \tag{7}$$

while the average shortest path has been shown to scale with $\ln N$ just like random networks [22].

Some small-world networks have been shown to exhibit another interesting property, along with high clustering and short path lengths. Networks such as the internet, collaboration networks, ecological networks, cellular networks, citation networks, and the community of actors exhibit power-law degree distributions [13].

Definition 5. A scale-free network obeys a power-law connectivity distribution. That is, the connectivity of a node is determined by a shape parameter and a scaling factor with the probability distribution function

$$f(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}, \quad k = 1, 2, \ldots, N, \tag{8}$$

where $\zeta(\gamma) = \sum_{x=1}^{N} 1/x^{\gamma}$ is the truncated Riemann ζ function, also called the scaling factor. The shape parameter of the distribution is $\gamma > 0$.

For scale-free networks, the usual range is $2 < \gamma \leq 3$, while $1 < \gamma < 2$ is typical for biological networks (genes, proteins, metabolism, and ecological networks) [13].

The power-law distributions exhibit a scale invariance. That is,

$$f(\beta \mathrm{k}) = \frac{(\beta k)^{-\gamma}}{\zeta(\gamma)} = (\beta)^{-\gamma} f(k), \tag{9}$$

where β is a constant. Because of this property, such networks have become known as scale-free networks.

In this paper we generate many Facebook group networks and compute their numerical measures in comparison to the values described above for random, small-world, and scale-free networks. With this in mind, we proceed to our discussion of network construction and analysis.

https://doi.org/10.25088/ComplexSystems.23.3.197

3. Network Creation Methodology

Facebook is arguably the most influential social network in human history [23]. With approximately 900 million users and over 125 billion friend connections, Facebook is available in more than 70 languages (these numbers were available on Facebook in March 2012) and is larger than any network before it, permeating every corner of the globe and every walk of life. Naturally, massive amounts of data facilitate a quest for knowledge and meaning. Several authors have managed to analyze Facebook as a network whose nodes are individuals and whose links are friendships [5, 6], identifying small-world properties. Those papers focus on individual users and analyze snapshots of the network at a certain point in time, thus focusing on the topological aspects of the network. Our work follows a similar approach at a Facebook group level. Other authors have focused on characteristics of information spread and information replication through the Facebook network [24] at the basic level of a "meme" (designating an idea or message that spreads and evolves analogously through communication) or in an aggregate fashion, inducing a coarsening of the network. In [24] the focus is on the dynamical evolution of the memes over time and a statistical assessment of the impact of mutations on the actual messages as they are replicated by users and friends. The authors use a genetic network approach. Although our work does not consider the dynamical aspects of Facebook, it does provide a new way of coarsening a Facebook subnetwork generated by common group interests. Our approach is to use "guided search" to find and analyze networks created from the latent data of the Facebook social graph that has interesting properties, for example, high clustering or short average path length. The size of Facebook, combined with privacy restrictions, poses a serious hindrance for attaining interesting data characteristic of an actual social network. In this section we describe a methodology for analyzing coherent, self-contained Facebook subnetworks of a tractable size.

The idea of analyzing Facebook groups is an intuitive continuation of the logic behind small-world networks. Groups are collections of people connected via a common context, a set of facts and circumstances that surround a situation, event, or concept. Naturally, to move beyond a set of isolated groups, it must be assumed that the average person belongs to multiple groups. This is feasible, as not many people would choose to define themselves by a single context. To this end, in order to analyze Facebook groups, we must operate under a small number of restrictions. First, the method for retrieving groups of a common context is limited to a search for groups with a single word in the group's title. For example, if we decided to form a network of groups with "word" in the title, we might have

https://doi.org/10.25088/ComplexSystems.23.3.197

F_{word} = {word, Words, I love words, WORD!!!, ...}. Furthermore, we can combine search results of multiple words to create multiword networks. To develop multiword networks as networks defined by a unifying context, we combined words that could define a community, circle of acquaintances, or field of interest. For example, to create a network from the words foo and bar, we might have F_{foo} = {**foo**, **FOO**D!, **foo**bar? No **Foo**ls, **FOO**BAR!!!,...} and F_{bar} = {**Bar** Hopping, **bar**istas unite, No Holds **Bar**red!!, FOO**BAR**!!!, foo**bar**?...}, from which we could form $F_{\text{foobar}} = F_{\text{foo}} \cup F_{\text{bar}}$.

Recall that two nodes of our group networks are adjacent or linked if they contain at least one mutual member. This procedure generates the edges. Once all the edges have been added, the networks are pruned as follows in order to reduce the skew of data: (1) all nodes with degree = 0 (isolated nodes) are removed; and (2) every node of every subgraph with average shortest path length $l = 1$ is removed.

We generate networks by using individual keywords as well as by combining multiple keywords to create larger networks with a unifying context for statistical purposes and to understand network growth and scaling. Not every word renders interesting results, but in general every network larger than approximately 40 nodes exhibited the same general degree distribution shapes, low average path lengths ($\approx$2), and high clustering, on average about three times the clustering coefficient for a random graph with the same number of nodes and edges.

4. Network Visualization and Analysis

This section contains the main results obtained by analyzing the Facebook group networks. We divide the section into two subsections that focus on network visualizations and numerical characteristics.

4.1 Network Visualization

In Figures 1 through 9 we start our analysis by providing a visualization of several networks. In each figure we generate the nodes of the network and the links between them. The visualizations are rendered via an implementation of the Fructerman–Reingold force-directed algorithm [25], which tends to create highly clustered "cores." We also provide the keywords used to generate the networks, as well as the numerical characteristics: the number of nodes N, the number of links L, the average connectivity $\langle k \rangle$, the average clustering coefficient C, and the average path length l. Observe the high density of the links in the networks anime, bieber, and muslim, for which we also provide an enlarged view of the core, as opposed to the very sparse army and

https://doi.org/10.25088/ComplexSystems.23.3.197

navy networks. We note as of now that the numerical characteristics of the anime, bieber, and muslim networks indicate small-world network features.

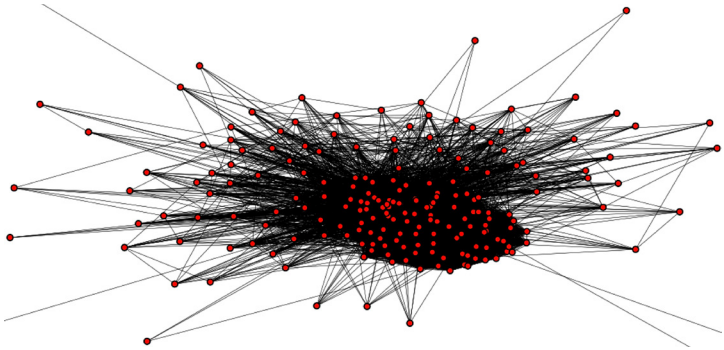

Figure 1. Facebook groups network created from groups with the word anime in the title. The numerical characteristics are $N = 276$, $L = 9770$, $\overline{C} = 0.708$, $l = 1.837$, $\langle k \rangle = 70.797$, indicating short path lengths and a clustering coefficient larger than the corresponding one for a random network with similar topological aspects, $\overline{C}_{\text{rand}} = 0.257$, $l_{\text{rand}} = 1.319$.

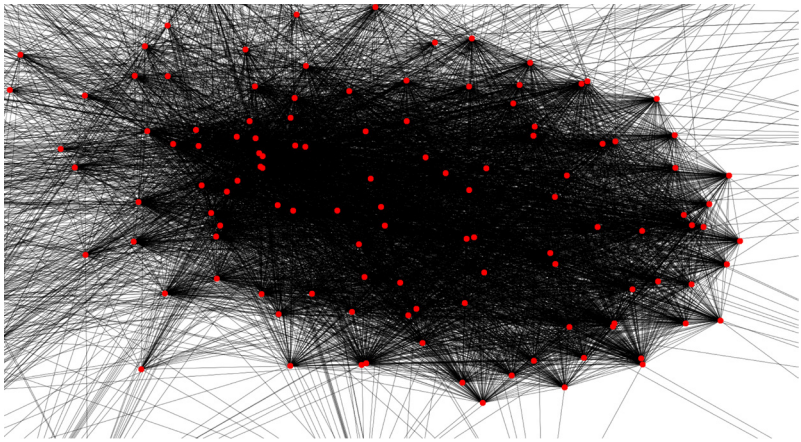

Figure 2. An enlarged depiction of the core of Figure 1.

https://doi.org/10.25088/ComplexSystems.23.3.197

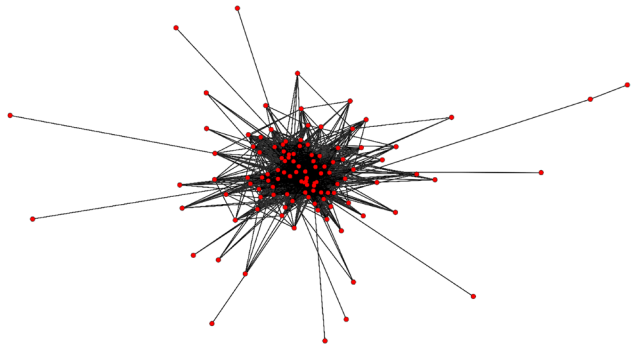

Figure 3. Plot of bieber network. The numerical characteristics are $N = 117$, $L = 1714$, $\overline{C} = 0.646$, $l = 1.903$, $\langle k \rangle = 29.299$, indicating short path lengths and a clustering coefficient larger than the corresponding one for a random network with similar topological aspects, $\overline{C}_{\text{rand}} = 0.25$, $l_{\text{rand}} = 1.41$.

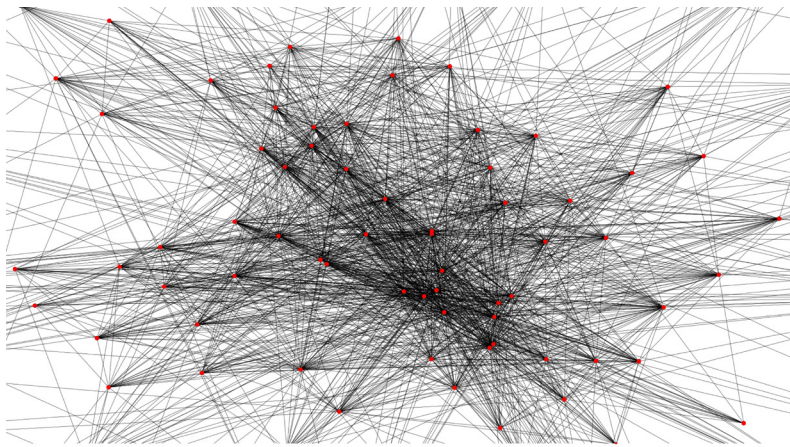

Figure 4. An enlarged depiction of the core of Figure 3.

https://doi.org/10.25088/ComplexSystems.23.3.197

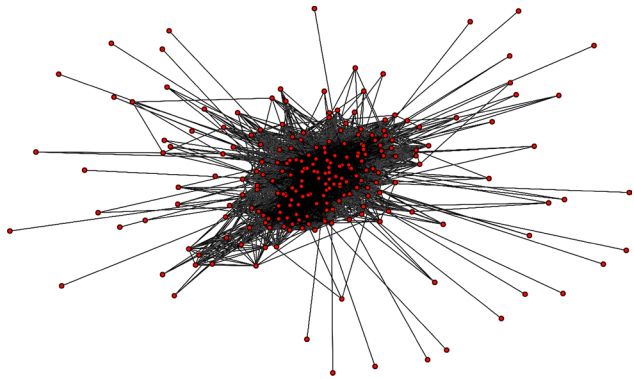

Figure 5. Plot of muslim network. The numerical characteristics are $N = 233$, $L = 4326$, $\overline{C} = 0.564$, $l = 2.121$, $\langle k \rangle = 37.133$, indicating short path lengths and a clustering coefficient larger than the corresponding one for a random network with similar topological aspects, $\overline{C}_{\text{rand}} = 0.159$, $l_{\text{rand}} = 1.508$.

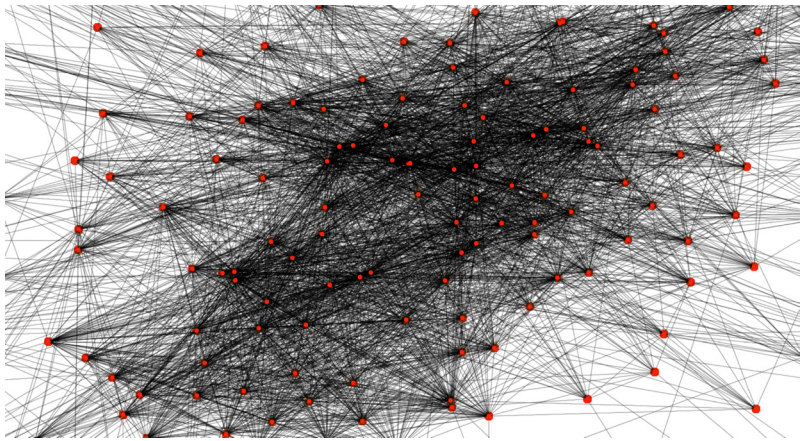

Figure 6. An enlarged depiction of the core of Figure 5.

Visualization is an important aspect to consider when analyzing networks. For example, the army network of Figure 7 may seem only slightly different from the navy network depicted in Figure 8. Yet, when visualized using the Fructerman–Reingold force-directed algorithm [25] (Figures 9 and 10), it becomes evident that while the army network is a single, moderately clustered network, the navy network is composed of two isolated subgraphs. The two main features of the Fructerman–Reingold force-directed algorithm are: (1) vertices con-

https://doi.org/10.25088/ComplexSystems.23.3.197

nected by an edge should be drawn near each other; and (2) vertices should not be drawn too close to each other. In general, how close vertices should be placed depends on how many there are and how much space is available. All vertices repel each other, but connected vertices attract. This leads to vertices with low connectivity being overcome by repelling forces and a core of highly connected vertices forming near the center of the graph. In this fashion, the model is driven, but not entirely constrained, by the physical analogy of a collection of charged particles connected by springs.

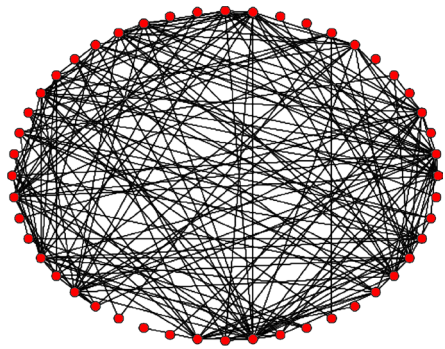

Figure 7. Plot of army network where the nodes are randomly distributed on a ring. The numerical characteristics are $N = 48$, $L = 189$, $\overline{C} = 0.434$, $l = 2.32$, $\langle k \rangle = 7.875$. It is difficult to gain much knowledge from a graph of this size, but it is useful when analyzing the behaviors of network parameters as N is increased.

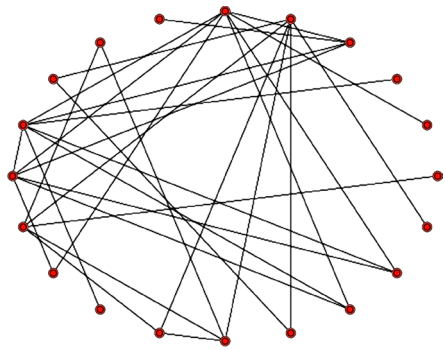

Figure 8. Plot of navy network where the nodes are randomly distributed on a ring. The numerical characteristics are $N = 20$, $L = 31$, $\overline{C} = 0.504$, $l = 1.83$, $\langle k \rangle = 3.1$. This is another example of a small network. It serves as a good conceptual comparison to Figure 7, due to its generating keyword.

https://doi.org/10.25088/ComplexSystems.23.3.197

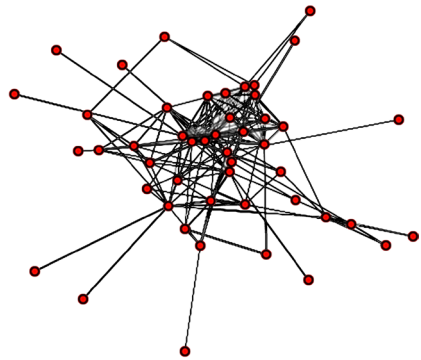

Figure 9. Plot of army network using the Fructerman–Reingold force-directed algorithm.

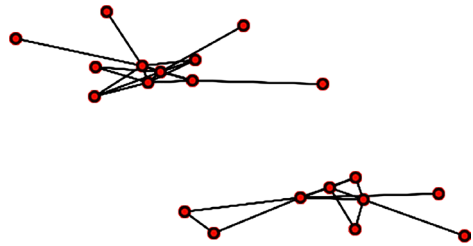

Figure 10. Plot of navy network using the Fructerman–Reingold force-directed algorithm.

Now, observe that the individual group networks presented in Figures 1 through 9 are fairly small. To make sure that our networks are large enough for a more accurate analysis, we have combined individual keywords into categories and then created larger networks from these categories. The categories that will be used in further graphical representations are presented in Table 1.

Figures 11 and 12 show what a larger network actually looks like and the complexity that manifests by doubling the number of nodes and adding more related groups. Other multiword graphs display comparable levels of complexity, but it should be noted that the three cores evident in Figures 11 and 12 are characteristic of the Religion multiword graph force-directed rendering and may or may not be seen in other graphs. The multiword graph Sports is presented in Figures 13 and 14 for comparison. Although the cores look slightly different in the Religion versus Sports visualizations, it is the mathematical analysis that will shed more light upon the similarities and differences between them. However, such visualizations allow us to have a picto-

https://doi.org/10.25088/ComplexSystems.23.3.197

rial view of the networks under consideration. We also notice the impact of increasing the network size on the level of visual complexity of the networks.

Category	Keywords
Religion	islam, muslim, bible, atheist, catholic, christian, god, allah, mormon, pagan, pope
Politics	republican, romney, vote, government, politic, election, democrat, obama, senate
Hobby	art, gun, fish, collect, craft, comic, cook, eat, hunt, stamp
Sexes	boy, man, gay, girl, lesbian, male, woman
Anime	naruto, bleach, dragonball, DBZ, OnePiece, Gundam, anime, saiyan, sasuke
Minorities	africa, white, black, american, china, asian, hispanic, latin, greek, KKK, nazi
Science	physics, fractal, chaos, nuclear, math, science
Music	bieber, gaga, beyonce, spears, kesha, kardashian, MTV
Sports	basketball, soccer, football, baseball, golf, tennis, bowling, swim
Computer	ipad, ipod, android, PC, mac, computer

Table 1. Ten categories of keywords used in the mathematical analysis of Facebook group networks.

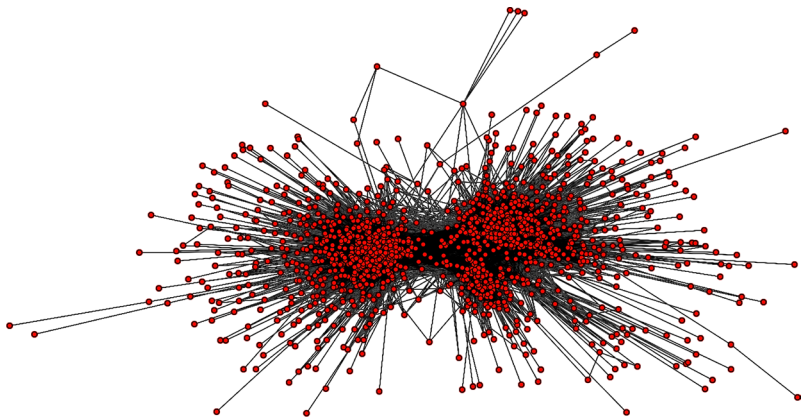

Figure 11. Plot of the Religion network. The individual keywords used are islam, muslim, bible, atheist, catholic, christian, god, allah, mormon, pagan, and pope. The numerical characteristics are $N = 1558$, $L = 39\,547$, $\overline{C} = 0.543$, $l = 2.723$, $\langle k \rangle = 50.766$.

https://doi.org/10.25088/ComplexSystems.23.3.197

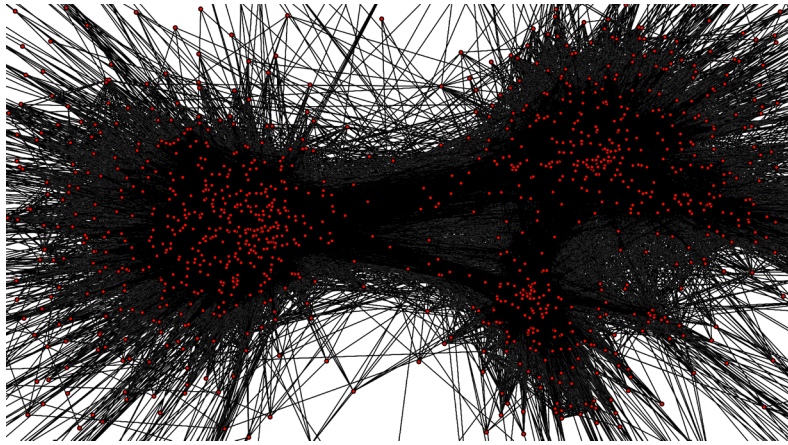

Figure 12. An enlarged depiction of the core of Figure 11. Observe the three smaller cores and notice the density of links.

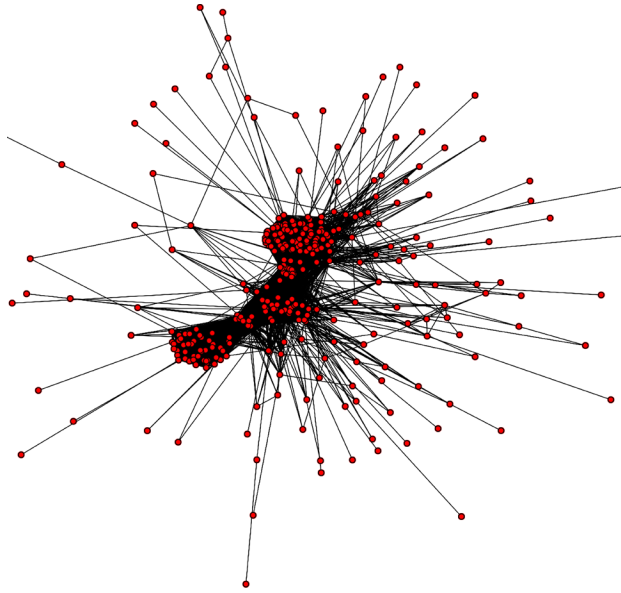

Figure 13. Plot of the Sports network. The individual keywords used are basketball, soccer, football, baseball, golf, tennis, bowling, and swim. The numerical characteristics are $N = 342$, $L = 10\,356$, $\overline{C} = 0.715$, $l = 2.403$, $\langle k \rangle = 60.561$.

https://doi.org/10.25088/ComplexSystems.23.3.197

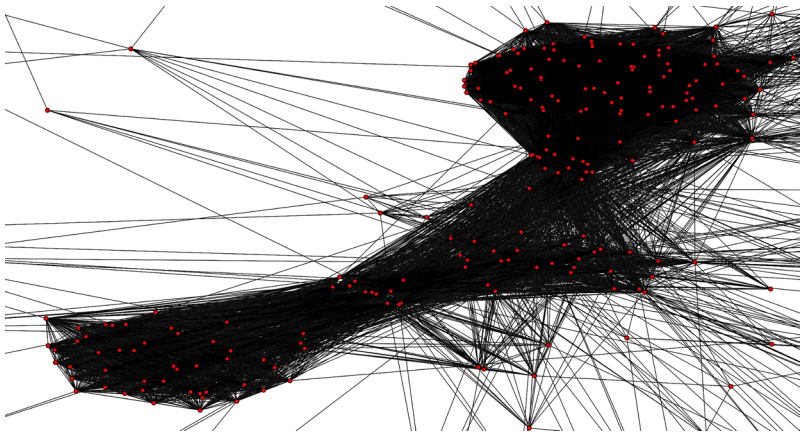

Figure 14. An enlarged depiction of the core of Figure 11.

Now that we have a pictorial view of some of the networks under consideration, we focus on the numerical characteristics and identify some mathematical similarities and differences between the categories. We observe that the studied networks exhibit average clustering coefficients three to 10 times larger than those for random networks, while maintaining relatively small values of the average path length. Thus we claim that both small-world properties 1 and 2 are satisfied for Facebook group networks. We provide details in what follows and refer to power-law distributions as well.

4.2 Numerical Characteristics

First let us look at the degree distribution for the networks of Table 1, indicated accordingly in the titles of Figure 15. For each network, we generate the distribution of the connectivity values k and we present the results on log-log plots. Observe that for smaller networks, the plots are not as structured. However, for larger networks, it is apparent that the distribution is decreasing with increased connectivity and has a mild curvature, which means that it appears to be somewhat linear, which would be typical for power-law distributions. We do not observe a pronounced peak typical for random networks. As a matter of fact, we have used the generalized Pareto fit tool of MATLAB [26] to fit a power-law distribution to several degree distributions. One sample is shown in Figure 16 for the Religion network, which is the largest in this analysis. The fit is not performed on a log-log scale, as seen from the graph. The parameters are γ, the power, and ζ, which is the truncated Riemann zeta function as described in equation (8). Observe that the best fit for all the data points yields $1 < \gamma < 2$, but very

https://doi.org/10.25088/ComplexSystems.23.3.197

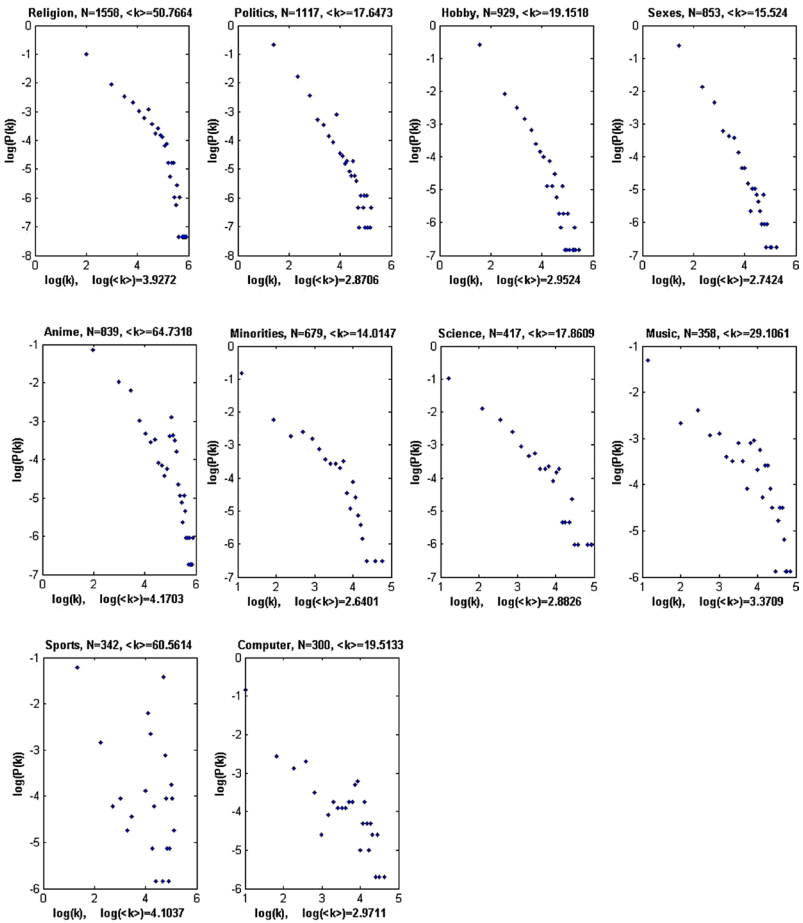

Figure 15. The distribution of the connectivity values in log-log plots for the 10 networks of Table 1, in decreasing order of network size. Each graph has the same general shape. As expected, the networks with more nodes render a log-log plot, as the plot appears to degrade at $N \approx 350$. Although the graphs appear to be mostly linear, the distributions cannot be characterized entirely by a power-law decay. In general, the plots adhere to a power law for small values of k but then seem to decrease more quickly than a power law would predict for large values of k. This is seen with more precision in Figure 16, where we attempt to fit a Pareto distribution to the degree distribution of the individual networks.

close to 1, and that the actual data tail seems to be lighter than the estimated distribution. Thus, if we were to fit only the tail, we would obtain larger γ values, which would be closer to what was found in the literature for various biological networks, for example. In Table 2

https://doi.org/10.25088/ComplexSystems.23.3.197

we present the fitted parameters for the largest networks. All of them yield similar results. Using a t distribution for small samples, we find a 95% confidence interval for the average γ to be (0.934, 1.499). Barabási and Albert provide parameter values for several networks with power-law degree distributions in [13]. It is interesting to note that the networks with similar values for γ ($\approx$1) are the Ythan estuary and Silwood park networks, both of which are undirected ecological networks. On the other hand, other internet networks have been shown to exhibit a power-law scaling factor $2 < \gamma < 3$ [13]. Recently, a study of the social graph of active Facebook users did not yield a strict power-law distribution for the degree distribution [5]. The same phenomenon is observed here for group networks, so a strict power-law fit may not be the most appropriate approach to Facebook degree distributions. But both the individual user network [5] and our aggregated network exhibit some similarities of the degree distributions, namely the monotonicity, the curvature, the fairly small degrees of typical users, or the rather large variance of degrees. From this perspective, the scaling of the Facebook network from users to groups formed as in this paper leads to invariant properties of the degree distributions. This holds mostly for aggregated networks of large enough size, like the first five graphs of Figure 15.

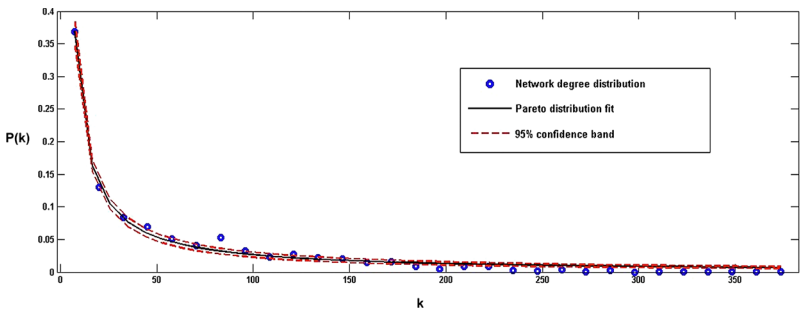

Figure 16. Pareto fit for the degree distribution. The estimated shape parameter γ is close to 1. However, the tail of the graph is actually lighter than the estimated fit. (N = 1558, $\langle k \rangle$ = 50.7664, ksi = 1 / 2.7523, and gamma = 1.0082).

At the same time, we provide a box plot representation of the values of the scaled connectivity values k / N for each of the 10 networks of Table 1, for an easier comparison of the connectivity with respect to the network size. This is shown in Figure 17, including the median, the first and third quartiles, and the outliers. Observe that only the Sports network exhibits a significant departure from the overall trend of small median, indicating small connectivity for most nodes. How-

https://doi.org/10.25088/ComplexSystems.23.3.197

ever, this is one of the smallest networks considered, and thus we need further analysis with larger networks under this category. We should mention here that both the ANOVA and Kruskal–Wallis tests for checking if the distributions represented in Figure 17 are the same yield p values close to zero even for fewer than 10 categories.

Keyword Category	N	ξ	γ
Anime	839	0.4889	0.9354
Sexes	853	0.2533	1.3875
Hobby	929	0.1922	1.4237
Politics	1117	0.3025	1.3258
Religion	1558	0.3633	1.0082

Table 2. Parameters of Pareto distribution, $p(x) = \xi(\gamma)\, x^{-\gamma}$, fitted to several multiple keyword networks.

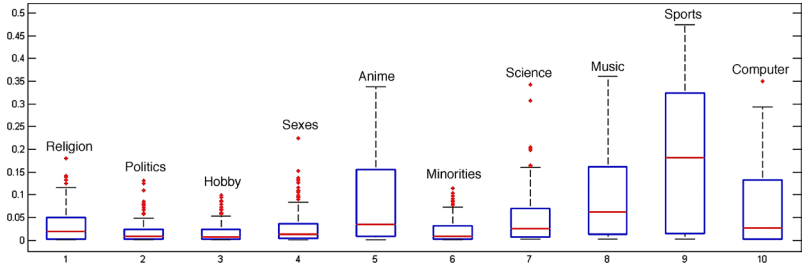

Figure 17. Box plot for the k/N values associated with the networks of Table 1.

Finally, in reference to the degree distribution, in Figure 18 we plot the average connectivity $Y = \text{Output} = \langle k \rangle$ as a function of $T = \text{Target} = N$ on a log-log plot. The plot clearly exhibits a positive correlation, and linear regression analysis yields $\log(\langle k \rangle) = 0.6 \log(N) + 0.19$, thus the average connectivity increases as a power of the network size, and slower than N, which is typical for random graphs. The correlation coefficient is $R \sim 0.78$, which indicates a moderate linear relationship between N and $\langle k \rangle$.

We conclude here that the degree distribution of group networks exhibits a monotonic decrease at a fairly steep rate with increased connectivity, partially fitted by a straight line, which corresponds to a power-law distribution. Most groups have a rather small connectivity in comparison to the size of the network. For smaller networks, like Music, Sports, and Computer, these results are not very accurate, so more analysis with an increased number of keywords in each category

https://doi.org/10.25088/ComplexSystems.23.3.197

will be needed in the future, paired with a more in-depth search for a best fit of the degree distribution.

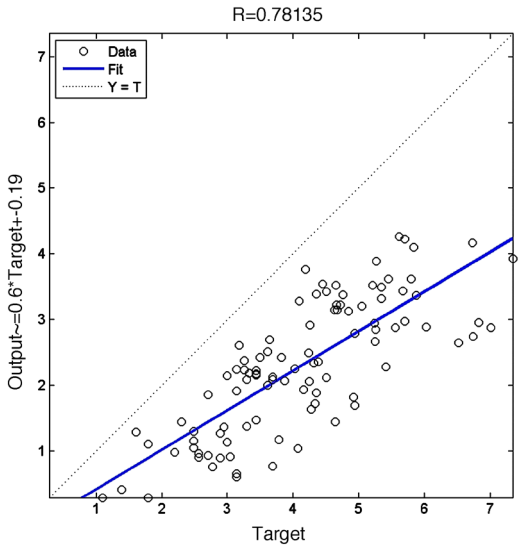

Figure 18. Log-log plot of $\langle k \rangle$ versus N for all the networks considered in this analysis, including single and multiple keywords. The fitted line shows $\langle k \rangle \propto N^{0.6}$.

Now let us look at the clustering coefficient defined in equation (2). We compute $\overline{C}(k)$ for all possible connectivity values k in each network and plot them against k on a log-log scale in Figure 19. What we find is similar to other small-world network studies [4, 5], where each plot exhibits an approximately logarithmic decay for large values of k.

Again we provide a box plot representation of the values of $C(k)$ for each of the 10 networks of Table 1 for an easier comparison of the average connectivity values for different degrees. This is shown in Figure 20. We note the differences between categories of networks, which are confirmed by both ANOVA and Kruskal–Wallis tests that lead to almost null p values. We note that overall the values are larger than the corresponding average clustering coefficients of random networks, thus indicating small-world networks. In fact, the average clustering coefficients for the networks listed in Table 1 are, on average, 14 times larger than the average clustering coefficients of random networks.

https://doi.org/10.25088/ComplexSystems.23.3.197

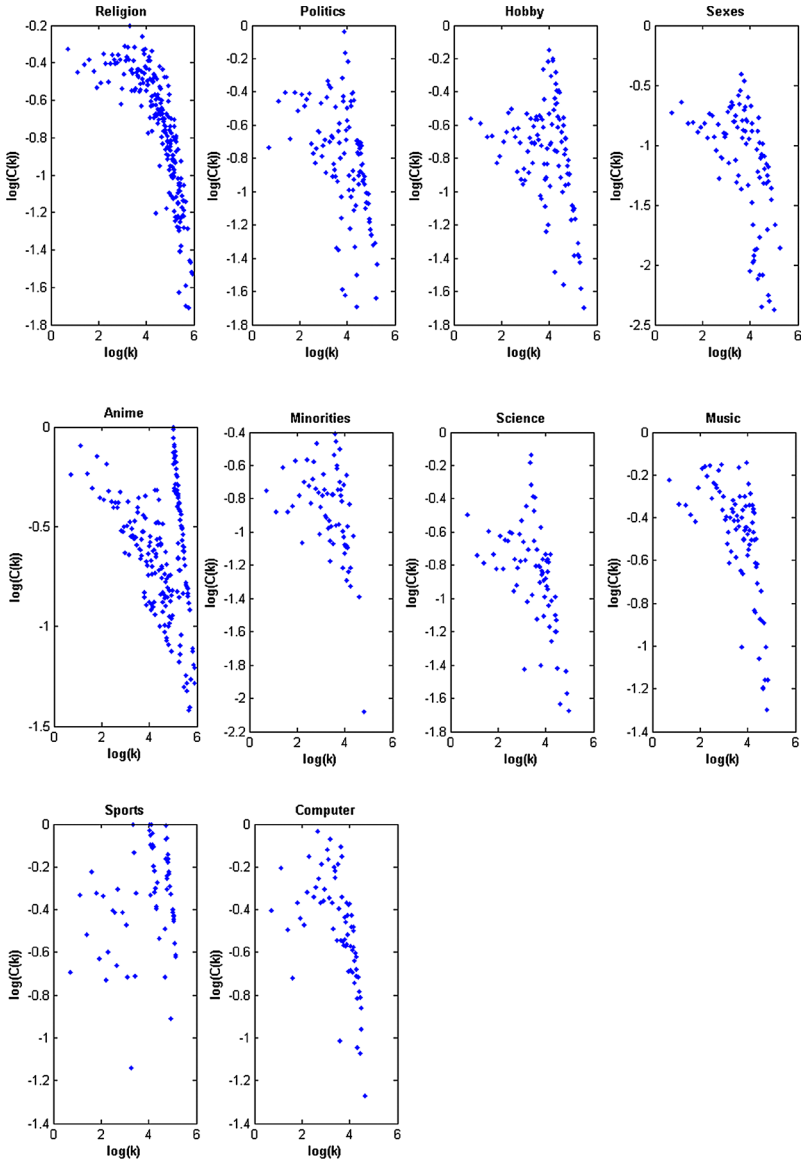

Figure 19. The average clustering coefficient $\overline{C}(k)$ against the degree k on a log-log scale for the 10 networks of Table 1, in decreasing order of network size. In general, each plot indicates an approximate logarithmic decay for large k. These results are similar to other studies of small-world networks [4, 5]. Observe that the largest network, Religion, provides the most orderly plot, which is to be expected.

https://doi.org/10.25088/ComplexSystems.23.3.197

Finally, in reference to the clustering coefficients, we present a plot of the average clustering coefficient $\overline{C}$ versus the network size N in Figure 21 (top graph, blue dots) for all the networks considered in this study, not only the 10 large networks constructed by aggregating smaller networks corresponding to the various categories. Most of the networks considered in this study have fewer than 400 nodes. We plot on the same graph the corresponding values $\overline{C}_{\text{rand}}$ for clustering coefficients of a random network with the same number of nodes and edges (top graph, red circles). We can notice immediately that the Facebook values are clearly higher than those for random networks, fulfilling property 2 of small-world networks. Not only that, but by focusing only on the Facebook data for $\overline{C}$ in Figure 22 we note an interesting result, as the Indian railway network has exhibited $\overline{C}$ values that remain relatively constant as the network size increases [4]. Figure 22 has too few data points to definitively decide on the behavior of $\overline{C}$ for large networks; however, to some extent it appears to steady out around $\overline{C} \approx 0.5$ for larger values of N. This suggests that we need more sample networks with a large number of nodes, which will be a subject for future research. At the same time, we will devise a more systematic procedure for increasing the network size. For example, adding keywords one by one for a slower increase of the network, besides expanding the selection of keywords and categories based on a dictionary search. These will be subjects for future research.

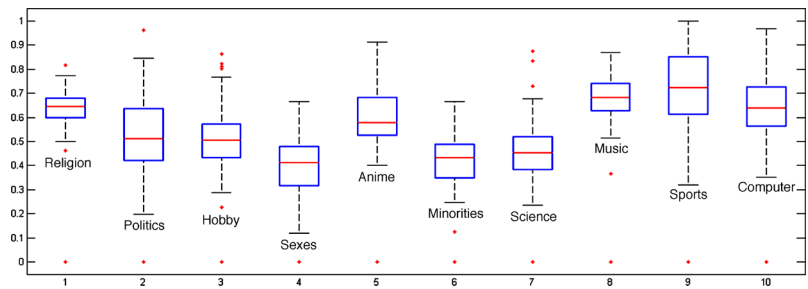

Figure 20. Box plot of the average clustering coefficients for individual connectivity values $C(k)$ associated with the networks of Table 1.

Now, regarding the average path length, in Figure 21 (bottom graph) we plot the average path length for both the Facebook networks of this study (blue dots) and the corresponding random networks (red circles). The plots are clustered around small values, and the average path length does not seem to exhibit an increase with increased network size. Notice the similarity between the Facebook

https://doi.org/10.25088/ComplexSystems.23.3.197

plots and random network plots, which suggests that property 1 of small-world networks is also satisfied. We also include a box plot of l, the average path length, as well as the corresponding distribution in Figure 23 for both Facebook networks and random networks. Notice that although the median path length for Facebook groups is higher than that of random networks, the variation is smaller, so that overall, the values for Facebook groups are slightly smaller than those for random networks. In general, the average path length is mostly less than three, which comes to supplement previous findings that the entire Facebook network of active users exhibits an average path length of about four [6], as opposed to the well-known six degrees of separation paradigm. Thus the Facebook group networks are indeed examples of small-world networks. In order to offer more precise statements regarding the distribution of average path length as shown in Figure 23 (right), we will have to generate a larger amount of data in the future. However, notice that despite the differences in the two distributions, they both have more or less the same behavior, which emphasizes property 1.

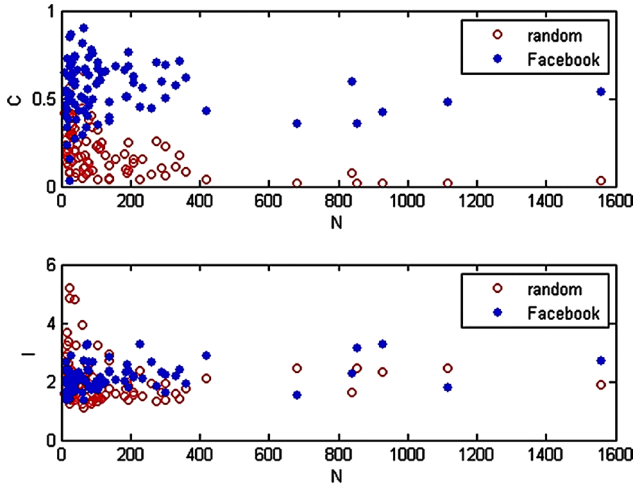

Figure 21. Average clustering coefficient (top) and average path length (bottom) as functions of N. The blue dots indicate the Facebook data and the red circles indicate the corresponding values for a random network as listed in Table 1. There is a clear difference for the clustering coefficients; however, it is hard to notice essential differences for the path lengths.

https://doi.org/10.25088/ComplexSystems.23.3.197

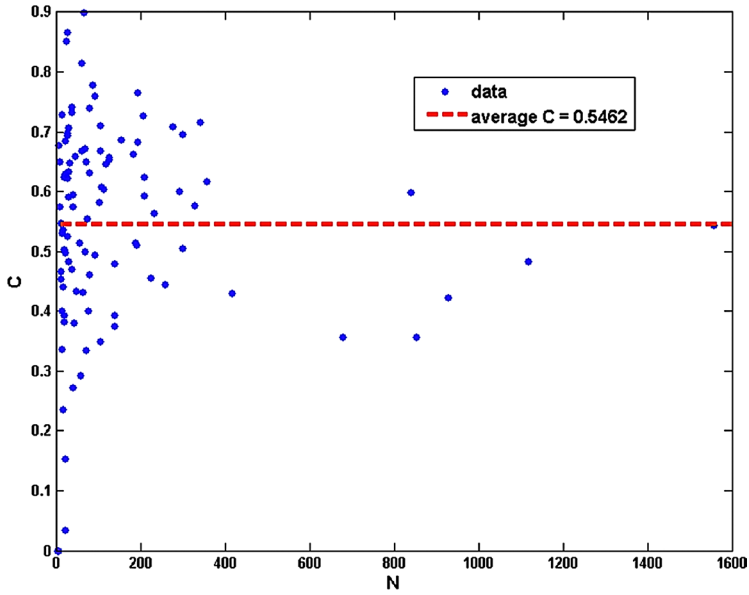

Figure 22. Average clustering coefficient as a function of N. The red dotted line depicts the average value and possible steady-state value for large N.

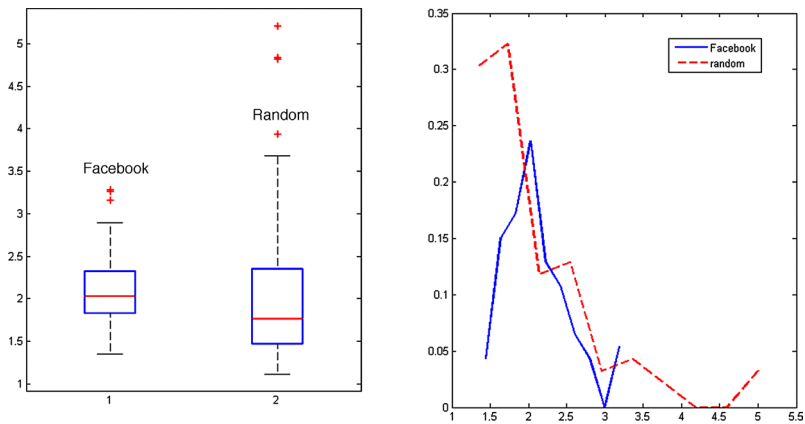

Figure 23. Average path length box plot (left) and distribution (right) for both the Facebook group networks and random networks with the same number of nodes and links.

We conclude that the Facebook group networks exhibit features of ordered networks and random networks. That is, they have large clustering coefficients like the ring lattice networks described by Watts

https://doi.org/10.25088/ComplexSystems.23.3.197

and Strogatz [9] that do not vanish with increased networks, and small average path lengths characteristic of random graphs [13]. Thus we conclude that the networks under consideration in this study exhibit small-world features. At the same time, the average connectivity increases as a power of the network size with approximation, while the average clustering coefficients and average path lengths do not exhibit a clear scaling with N.

In the future we plan to expand this analysis to more and larger networks and to supplement it with a more in-depth study of the numerical characteristics presented in this paper and other suitable measures.

5. Conclusions and Future Work

In this paper we provided a study of the structure of Facebook group networks. We generated a number of networks using keywords for group selection, and we linked groups with common members. We focused on a number of measures that can describe the structure of these networks. Our networks have degree distributions that cannot be entirely characterized by a power law, clustering coefficients that are significantly larger than what would be expected for random networks, and consistently small values for the average shortest paths, characteristic of random graphs. That is to say, our analysis has shown that Facebook group networks are small-world networks. A unique element of this study is that we did more than analyze only one network or a handful of networks; we analyzed an ensemble of real-world networks and were able to find relationships that otherwise could not have been found. Figures 18 and 22 attest to this fact and suggest an interesting path to take in future studies.

Although our methodology for network construction is efficient, it is not the only possibility. We chose keywords that would most likely render interesting and mostly unambiguous results using a programmatic approach. Our future work will expand this analysis to include more networks with larger sizes, in order to be able to generate a more complete statistical analysis of Facebook group networks. Future studies should include other measures of interest such as efficiency, correlations, or spectral properties, as well as more sophisticated statistical analyses.

Moreover, it would be important to analyze the dynamics of the networks over time, since Facebook groups could be added or removed from the networks, based on the natural changes that occur, and links readjusted. This would imply observing Facebook over a significant amount of time to have sufficient data for statistical purposes.

https://doi.org/10.25088/ComplexSystems.23.3.197

At the same time, it would be of interest to generate a dynamical system model of the social networks described here. More precisely, we can define some meaningful states of the nodes and associated rules that could help us understand the complexity of this kind of large-scale network. For instance, by focusing on the links in the networks, which are created based on individuals who are common to multiple groups, we could put a weight on the influence of one group over another group (the target) in terms of Facebook messages that are posted in the target group by the common members during a unit of time. By defining a suitable threshold function, a node could be labeled as active or inactive, based on the aggregated influences of the individual input nodes through the common members. This could generate a Boolean network whose dynamics could be studied to provide some insight on the impact of multiple memberships of individuals on the activity of groups.

References

[1] R. Albert, H. Jeong, and A.-L. Barabási, "Internet: Diameter of the World-Wide Web," *Nature*, **401**, 1999 pp. 130–131. doi:10.1038/43601.

[2] T. Helikar, J. Konvalina, J. Heidel, and J. A. Rogers, "Emergent Decision-Making in Biological Signal Transduction Networks," *Proceedings of the National Academy of Sciences*, **105**(6), 2008 pp. 1913–1918. doi:10.1073/pnas.0705088105.

[3] V. Latora and M. Marchiori, "Is the Boston Subway a Small-World Network?," *Physica A: Statistical Mechanics and Its Applications*, **314**(1–4), 2002 pp. 109–113. doi:10.1016/S0378-4371(02)01089-0.

[4] P. Sen, S. Dasgupta, A. Chatterjee, P. A. Sreeram, G. Mukherjee, and S. S. Manna, "Small-World Properties of the Indian Railway Network," *Physical Review E*, **67**(3), 2003 p. 036106. doi:10.1103/PhysRevE.67.036106.

[5] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The Anatomy of the Facebook Social Graph." arxiv.org/abs/1111.4503.

[6] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna, "Four Degrees of Separation." arxiv.org/abs/1111.4570.

[7] S. Milgram, "The Small World Problem," *Psychology Today*, **2**, 1967 pp. 60–67.

[8] J. Travers and S. Milgram, "An Experimental Study of the Small World Problem," *Sociometry*, **32**(4), 1969 pp. 425–443.

[9] D. J. Watts and S. H. Strogatz, "Collective Dynamics of 'Small-World' Networks," *Nature*, **393**, 1998 pp. 440–442. doi:10.1038/30918.

[10] P. Erdős and A. Rényi, "On Random Graphs," *Publicationes Mathematicae Debrecen*, **6**, 1959 pp. 290–297.

https://doi.org/10.25088/ComplexSystems.23.3.197

[11] R. Serra, M. Villani, and L. Agostini, "A Small-World Network Where All Nodes Have the Same Connectivity, with Application to the Dynamics of Boolean Interacting Automata," *Complex Systems*, **15**(2), 2004 pp. 137–155. http://www.complex-systems.com/pdf/15-2-3.pdf.

[12] C. Aguirre, F. Corbacho, and R. Huerta, "Static and Dynamic Properties of Small-World Connection Topologies Based on Transit-Stub Networks," *Complex Systems*, **14**(1), 2003 pp. 1–28. http://www.complex-systems.com/pdf/14-1-1.pdf.

[13] R. Albert and A.-L. Barabási, "Statistical Mechanics of Complex Networks," *Reviews of Modern Physics*, **74**(1), 2002 pp. 47–97. doi:10.1103/RevModPhys.74.47.

[14] D. J. Watts, *Six Degrees: The Science of a Connected Age*, New York: W. W. Norton & Company, 2003.

[15] A.-L. Barabási, *Bursts: The Hidden Pattern behind Everything We Do*, New York: Dutton, 2010.

[16] A.-L. Barabási, *Linked: How Everything Is Connected to Everything Else and What It Means*, New York: Plume, 2003.

[17] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, "Classes of Small-World Networks," *Proceedings of the National Academy of Sciences*, **97**(21), 2000 pp. 11149–11152. doi:10.1073/pnas.200327197.

[18] H. De Sterck, V. E. Henson, and G. Sanders, "Multilevel Aggregation Methods for Small-World Graphs with Application to Random-Walk Ranking," *Computing and Informatics*, **30**(2), 2011 pp. 225–246.

[19] M. E. J. Newman, "The Mathematics of Networks," in *The New Palgrave Dictionary of Economics*, 2nd ed. (L. E. Blume and S. N. Durlauf, eds.), Basingstoke, UK: Palgrave Macmillan, 2008 pp. 1–12.

[20] R. Cont and E. Tanimura, "Small-World Graphs: Characterization and Alternative Constructions," *Advances in Applied Probability*, **40**, 2008 pp. 939–965. doi:10.1239/aap/1231340159.

[21] D. J. Watts, "Networks, Dynamics, and the Small-World Phenomenon," *The American Journal of Sociology*, **105**(2), 1999 pp. 493–527.

[22] M. E. J. Newman, *Networks: An Introduction*, New York: Oxford University Press, 2010.

[23] P. Adams, *Grouped*, Berkeley: New Riders, 2012.

[24] L. A. Adamic, T. M. Lento, E. Adar, and P. C. Ng, "Information Evolution in Social Networks." arxiv.org/abs/1402.6792.

[25] T. M. J. Fruchterman and E. M. Reingold, "Graph Drawing by Force-Directed Placement," *Software—Practice and Experience*, **21**, 1991 pp. 1129–1164.

[26] *MATLAB*, Release Version 7.10.0 (R2010a), Natick, MA: The MathWorks Inc., 2010.

https://doi.org/10.25088/ComplexSystems.23.3.197