Markov Stability Analysis and Community Structure in Social Networks

K Nikaj and M Ifti
Department of Physics, Faculty of Natural Science, University of Tirana, Boulevard Zogu I No.25, 1001 Tirana, Albania
E-mail: klotildanikaj@hotmail.com

Abstract. Individuals connected to realistic networks exhibit collective behavior. In order to characterize this phenomenon and explore the correlation between collective behaviors and locally interacting elements, we use statistical methods and visualization software as a combined approach to understand the behavior of the network for a given behavior of the agents that we use to recreate our network. The aim of this work is to identify the communities as hierarchical structures trying to find them between a giant component and a small-world network. By analyzing the data and describing how these networks fall in community structure, we aim to obtain new tools and methodology which will help us to describe how networks grow and fall apart in smaller structures, which have similar features with the large network, but different dynamics.

1. Introduction
Community detection is a popular topic, and many methods have been proposed to solve this problem, some of which use the notion of distance [1], measured on the set of graph nodes. When we talk about community and social networks, we need to define some terminology [1–4]:

- **Agents**, also called nodes or vertexes, refer to an individual that can have a relationship with other individuals.
- **Relation**, also called ties or edges, describes a relationship between two agents.
- **Network**, also called a graph, refers to a collection of agents and ties between them.
- **Community**, is a connected sub graph of a network in which the internal density of edges is significantly larger than the external one.
- **Degree Distribution**, is the distribution of the number of nodes with each degree value [1,12].
- **Clustering Coefficient**, is the degree to which nodes in a large social network tend to cluster together [1,12].

Network data (also referred to as social network data, real graph data) have attracted attention, and understanding patterns in this data has become an important research problem with many existing models [2,4–6]. Several previous studies suggest that properties of the entire network are quite different from the properties at the community level [2], so we can find many interesting features if we focus on the analysis of the community structure. When members of one community interact with members of another community by exchanging information [8–10] and ideas, at the same time they are leading to reactions between community members. Networks...
Table 1. Features of the real and BAM networks.

|                | Number of nodes | Number of links | Degree distribution | Clustering Coefficient |
|----------------|-----------------|-----------------|---------------------|------------------------|
| Real network   | 4039            | 88234           | 43.7                | 0.591                  |
| BAM network    | 4500            | 98324           | 5.6                 | 0.012                  |

with communities appear to have similar degree distribution and topologically interesting entities [2], but also distinct dynamical processes at community levels. In summary, we contribute to this work by exploring a vital connection between the persistence probability of a community at the stationary distribution and its local topology. Indeed, the research results show a profound difference in the topological properties of networks [2]. To verify our work, we use both, synthesized data with embedded communities and real-world social network data, by simulating the topology and dynamics of a community network.

2. The proposed method

Barabasi and Albert [1,11,12] have given the first explanation of the scale-free distribution, by reformulating Simon’s model [6] in the context of growing networks. In real networks, every instant new nodes and links are created and removed. The network grows as in the Barabasi Albert Model (BAM). In each instant, a new node is added to the network and is connected to an old node with a probability depending only on the degree of the target. The model includes two parameters: the number of the new agents and the availability (i.e. probability) of the old agents to collaborate. In each time step, a new group is formed, and the members that join it could be new or old agents. In this work, we propose a BAM modification, focusing on the interaction of agents in small groups, which can lead to changes in the structure of the network over each time step. We start by comparing some topological features of the BAM and those of a real network taken from an open database [7]. Tab.1 shows the data that we take into consideration for this work.

By comparing the clustering coefficient for the two networks, as we can see, the real network is almost 50 times larger than that of the BA network of the same size (precisely 46.1-times larger). As we see from Tab. 1, the degree distribution is different for the two networks, but if we compare even other topological features we will notice different values [11,12,15].

2.1. Analysis of the degree distribution

Let us suppose that $q$ is the probability that an agent is part of a community, then $1 - q$ is the probability to be not a member of this community. Assuming the same probability $1 - q$ for each agent, if an agent has a degree distribution $d$, $(1 - q)^d$ is the probability that agents are not in direct contact with each other. So we can write,

$$1 - q = \sum_d (1 - q)^d P(d),$$

where $P(d)$ is the probability of the agent to have $d$ neighbors, or the degree distribution. The probability of agent (A) having $d$ acquaintances in common with randomly chosen agent (B) is \( \binom{n}{d} p_d q_0^{n-d} \), where $p_d q_0^{n-d}$ is the probability that A knows B and each of his remaining $(n-k)$ acquaintances do not know B. The binomial coefficient is to include that there are \( \binom{n}{d} \) ways of selecting these $d$ from the $n$ people whom A knows. Let us randomly choose an edge among the \( \binom{n}{k} \) possible edges. After this, let us choose an other edge among the remaining $\binom{n}{d}-1$, and let us continue this process so that if already $k$ edges are fixed, any of the remaining $\binom{n}{d}-k$ have
equal probabilities to be chosen as the next one. The mean value of this binomial distribution is \( np_0 \) and the variance is \( np_0q_0 \).

A more realistic model can be obtained assuming that relations are ruled by a Markov Chain process [11] therein:

\[
\text{Prob}(B \in K_{Ak} | B \in K_{Ak-1}, \ldots, B \in K_{A_1}) = \text{Prob}(K_{Ak} | K_{Ak-1}) = b
\]

where \( b \) is a constant to be statistically estimated. Let \( K_{Ak} \) be the number of acquaintances of \( A' \)s and \( A_1, A_2, \ldots, A_n \) the individuals of the community. So we can write,

\[
\text{Prob}(K_{An} | K_{An-1}, \ldots, K_{A_1}) = \text{Prob}(K_{A_1}) b^{n-1} = \left(1 - \frac{n}{N}\right) b^{n-1}.
\]

If we choose \( k = 2 \) we obtain,

\[
\text{Prob}(K_{A_1} | K_{A_2}) = \left(1 - \frac{n}{N}\right) b = 1 - \frac{2n}{N} - \frac{m_2}{N},
\]

and \( b \) can be calculated as,

\[
b = \frac{1 - \frac{2n}{N} - \frac{m_2}{N}}{1 - \frac{n}{N}}.
\]

Here we chose \( m_2 = 10 \), where \( m_2 \) is chosen randomly and represents the number of acquaintances in common to two people, \( n = 10^3, N = 10^6 \), then \( b = 0.9999, p_0 = \frac{n}{N} = 0.001 \) as before and

\[
p_{12} = (1 - p_0) \left[ 1 - \left(1 - \frac{n}{N}\right) b^{n-1} \right] = 0.095
\]

\[
p_2 = (1 - p_0)(1 - p_1)p_2'
\]

where \( p_{12} \) denotes the joint probability that \( A \) and \( B \) do not know each other, and they have no common friends, and that \( A \) has at least one friend whom knows at least one friend of \( B \); \( p_2 \) denotes the probability that \( A \) and \( B \) do not know each other, but have at least one acquaintance in common; \( p_2' \) is the probability of different people who are the friends of acquaintances of the \( n \) people whom \( A \) knows and can be written as

\[
p_2' = 1 - \left(1 - \frac{n}{N}\right) b^{n-1}.
\]

These calculations still do not completely fit the data. The proposed model (Barabási Albert modified model) is applied to generate a social network of 4500 nodes (almost the same number of nodes as the real data set). We consider the probability \( p \) of creating a new node linked to another node equal to 0.001 as Eq. 6. A total of 98254 links are created by the model. We assume that the probability distribution was a Markov Chain function (Eq. 2), where the variation of the centrality measures the asymptotic probability for a random walker on the network to be standing on node \( i \) or jumps to any node. In each step of the Markov Chain we assume a probability \( q \) that the walker moves to a randomly chosen neighbor.

3. Simulation results and discussion

3.1. Data sets Information

The data sets are gathered from publicly available data, by accessing the public web interface provided by the sites. In our case, the data set is taken from the site of the Stanford University (see http://snap.stanford.edu/) [7] composed by an anonymous edge list from Facebook, to
verify the result of our study. This network is composed of 4039 nodes and 88234 edges. The average clustering coefficient of the network is 0.6 and the diameter of the network is 8. These statistical properties are calculated by using GepHi (open graph visualization platform).

We develop an approximation to the stability analysis based on an averaging Lyapunov criterion [4]. The best way to evaluate our approximated model is to validate it on real-world networks with known community structures.

Although synthesized networks might not reflect all the statistical properties of real ones, they can provide us with the known ground truths via planted communities and the ability to vary other network parameters such as sizes, densities and overlapping levels, etc. Fig. 1 shows randomly distributed nodes (left), and (right) the dense nodes (in other words the community structure). Using the modified BAM to generate the network, new agents belonging to a new cluster are mapped with different colors as shown in Fig. 3. The model captures two basic features of networks that can influence by varying the values of $p(1)$ and $q(2)$:

- the distribution of the connection type between agents, as shown in the Fig. 3;
- the overall connectivity of the network.

![Figure 1](image1.png)
**Figure 1.** Two snapshots of the real data set of the network using GepHi: A) randomly distributed nodes; B) nodes unfolding of communities [14].

![Figure 2](image2.png)
**Figure 2.** Growth process in the modified model: link creation between nodes (left), small cluster creation for small $p$ (middle), and small clusters tempt to join the giant component if we keep increasing $p$ (right).

Fig. 4 shows the percentage of agents belonging to the giant component, which is the largest connected chain of small clusters. The structure of the network can change, obtaining different emergent typologies. At each step, a new cluster composed from at least four nodes is created; for small $p$ the network is formed from small clusters, as $p$ increases the largest clusters start
Figure 3. a) An overabundance of new agents linked together, shows the possibility of creating a new cluster or community without taking advantage of an existing one. b) and c) show the probability of interaction between old agents and new ones, and the interaction between two old agents. d) shows agents belonging to multiple clusters at the same time, indicating a lack of diversity.

Figure 4. Fraction of agents belonging to the giant component.

Figure 5. Fraction of clusters composed at least from 4 nodes (a) and the number of communities versus the number of the users (b).

to have loops and become a densely connected cluster. The model displays a continuous phase transition, as the fraction $p$ increases from 0 to one. The distribution of different types of links reflects the different role of the agents. Indeed, if we have repeated cooperation between agents,
it is less likely that they have innovative ideas because they tend to have the same knowledge. In contrast, communities with various types of links are likely to have a more diverse perspective.

The network is reconstructed following the same dynamics observed when it comes apart under attack. If we want to destroy the network, we will start by removing the most connected nodes, and the giant component will be divided in more and more small clusters [1,12]. Now we started by creating small clusters, and the most connected node of each cluster tends to follow highly connected nodes, leading to the recreation of a giant component network. Fig. 4 shows a continuous transition, where all the most connected nodes of the small clusters join all the other most connected nodes, and by joining with each other, a large connected chain composed by all the small clusters is formed.

Fig. 5 (a) indicates the correlation between the clustering coefficient [13,15] and the number of clusters composed at least from four nodes. The correlation has an important influence on the topological properties of networks and can impact-related problems on networks, such as dynamic processes, or spreading information [4, 6, 10, 11]. Fig. 5 (b) shows the number of communities extracted versus the number of users [14]. As the figure shows, there are a few agents, with an enormous degree, and the majority of them have less than 100 connection. Our network can be categorized as a scale-free network. According to this assumption, as the network grows, new clusters of nodes are likely to connect to high degree nodes, since new agents tend to follow important agents (i.e., highly connected nodes). We can conclude that agents tend to form clusters according to acquaintances links, then these clusters tend to join the giant component of the network (see Fig. 4), following a phase transition of the network for different probabilities.

4. Conclusions

In this study we combine the statistical methods with the visualization software to study the network behavior under a controlled behavior of the network agents. The most important result is that we can reconstruct the network following the same dynamics observed when it comes apart under attack. We concluded that agents tend to form clusters according to acquaintances links, then these clusters tend to join the giant component of the network, following a continuous phase transition of the network for different probabilities. Social networks are evolving into major marketing tools in different areas. We can conclude that this type of combined research is very meaningful and its future applications may be able to provide decision support for public opinion monitoring, marketing, etc.

References

[1] Barabasi A L and Albert R (1999) Science 286 5439 pp. 509-512
[2] Pasta M Q and Zaidi F (2017) IEEE Access 5 2169-3536
[3] Jia P, MirTabatabaei A, Friedkin N E, Bullo F (2015) SIAM Review 57 3 367-397
[4] Aparicio S, Villazn-Terrazas J and Ivarez G (2015) Entropy 17 5848-5867
[5] Foss S, Shnerr S, Turylikov A (2012) Stoch. Syst. 2 1 208-231
[6] Simon H A (1995) Biometrika 286 425-440
[7] http://snap.stanford.edu/
[8] Iriarren J L, Moro E (2009) Phys. Rev. Lett. 103 038702
[9] Ver Steeg G, Galstyan A (2013) In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining Rome Italy 4-8 February 2013 pp. 3-12
[10] Kim M, Newth D, Christen P (2014) In Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion Seoul, Korea, 7-11 April 2014 pp. 753-758
[11] Newman M E J (2003) SIAM Review 45 167-256
[12] Albert R, Barabasi A L (2002) Rev. Mod. Phys. 29 47-97
[13] Latapy M (2008) Theo Com. Sci. 407 1-3
[14] Blondel V D, Guillaume J L, Lambiotte R, Lefebvre E (2008) J. Stat. Mech. P10008
[15] Lambiotte R, Delvenne J C, Barahona M (2015) IEEE Transactions on Network Science and Engineering 1 76-90