COMMUNITY DETECTION IN COMPLEX NETWORKS VIA DYNAMICS

Marcos Maia
Laboratório Associado de Computação e Matemática Aplicada
Instituto Nacional de Pesquisas Espaciais
Brazil
mdanielnm@gmail.com

Elbert Macau
Laboratório Associado de Computação e Matemática Aplicada
Instituto Nacional de Pesquisas Espaciais
Brazil
elbert.macau@inpe.br

Abstract
We use the dynamics of complex networks to identify communities. A well appropriated mathematical approach is used to study the behavior of the model's solutions. A quality function called clustering density is introduced to measure the effectiveness of the communities identification. Illustrations with real networks with community structure are presented.

Key words
Complex networks, Communities, Dynamics

1 Introduction

The last two decades has witnessed a growing interest on the study of complex networks and its properties [Chung and Lu, 2006]. The complex networks are around us, and us as individuals, are units of several social networks as family, work-groups, brotherhoods and many others [Boccaletti, 2006]. Such networks are mathematically modeled as graphs, which is composed of a finite set of vertices and a finite set of links.

Although commonly complex, real-world networks usually have a certain level of organization. For instance, its quite common having the links distribution on the graph frequently heterogeneous with high concentration of links inside specifics groups of vertices whereas there is a low concentration of vertices between those groups. This feature of complex networks is known as community structure [Fortunato, 2010]. Actually, there is no a precise definition of this property, nevertheless, we will use the following one:

Definition 1 (Community). *A community in a graph (or network) is a subset of this graph that has density of inside links larger than 0.5.*

The Figure 1 brings an example of a network with communities structure.

There are several classical methods used to detect communities and the typical algorithms include those

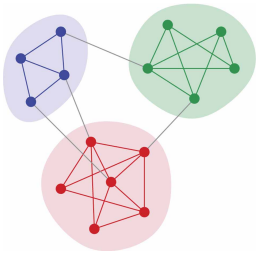

Figure 1. An example of a network with community structure. Credit: [Newman, 2012].

designed to maximize the value of the function known as "modularity", and those based on the topology structures directly, e.g., betweeness, degree, clustering coefficient and others. Many researchers have showed that the topological scales of complex networks can be reveled by their dynamics [Arenas et al, 2006; Moreno et al, 2004; Oh et al, 2005; Zhou et al, 2007]. A detailed survey on the problem of community detection can be found in [Fortunato, 2010].

Based on the works of [Wu et al, 2012] and [Nishikawa et al, 2004], we introduce here a new methodology for community detection. It is based on a Generalized Kuramoto Model (GKM) with a Fourier Term (GKMF) as the dynamics for the oscillators. We do not use the typical computer-made GN and LFR networks to test the model (and algorithm) to find communities because neither of theses models, GKM and GKMF, works well in large complex networks, mainly because its solutions are limited to the unit circle $\mathbb{S}^1 \subset \mathbb{R}^2$ and its quite common for different communities merge their temporal series. Our aim is to show that the networks dynamics can, indeed, be a new tool for community detection algorithms.

2 The Model and a Little bit More

Let $\mathcal{G}$ be a simple, non-directed and connected graph with n vertices. The adjacency matrix $A = [A_{ij}]$,

$i, j = 1 \cdots, n$ is a matrix that encodes the topological information of $\mathcal{G}$, i.e., $A_{ij} = 1$, if and only if there is a link between the vertices i and j, and $A_{ij} = 0$ otherwise. The degree g_i of a vertex i is the number of connections its receive. It can be written in terms of the adjacency matrix as $g_i = \sum_{j=1}^{n} A_{ij}$.

For each fixed vertex i in $\mathcal{G}$, the generalized Kuramoto model (GKM) reads:

$$\dot{\theta}_i = \omega_i + \frac{K_P}{n} \sum_{j=1}^{n} A_{ij} \sin(\theta_j - \theta_i) + \frac{K_N}{n} \sum_{j=1}^{n} (1 - A_{ij}) \sin(\theta_j - \theta_i), \quad (1)$$

$i = 1, \cdots, n$, where the ω_i is the natural frequency of vertex i, $K_P > 0$ and $K_N < 0$ are the coupling parameters. Note that the role of K_P is to attract those vertices that are connected ($A_{ij} = 1$) whereas the role of K_N is to push those vertices that are not connected ($A_{ij} = 0$). Note also that each trajectory $\theta_i(t)$ oscillates over the unit circle $\mathbb{S}^1 \subset \mathbb{R}^2$ and therefore we use the equivalence class $\mod 2\pi$ in all numerical integration.

The Theorem 1 give us an idea on how to choose the parameters K_P and K_N.

Theorem 1. *Let $\langle g \rangle = \sum_{i=1}^{n} g_i$ be the mean degree of a graph. If the natural frequencies ω_i of Eq.* (1) *are taken over a normalized distribution with* 0 *(zero) mean and finite variance and the solutions of Eq.* (1) *converge to fixed points then*

$$|K_N| \leq \frac{K_P \langle g \rangle}{(n - \langle g \rangle)}. \quad (2)$$

For now on, we will present and use our *Generalized Kuramoto Model with Fourier term* (GKMF). Let $\omega_i = \omega$ for all $i = 1, \cdots, n$ and consider the variable transformation $\theta_i(t) \to \theta_i(t) + \omega t$. The GKMF reads:

$$\dot{\theta}_i = \frac{K_P}{n} \sum_{j=1}^{n} A_{ij} \sin(\theta_j - \theta_i) + \frac{K_N}{n} \sum_{j=1}^{n} (1 - A_{ij}) \sin(\theta_j - \theta_i) + \frac{K_F}{n} \sum_{i=1}^{n} \sin[\rho(\theta_j - \theta_i)] \quad (3)$$

where $K_F > 0$ is the strength of the Fourier term and $\rho \in \mathbb{N} \setminus \{0, 1\}$ is the order of this term. As we will see, the role of the Fourier term is to force the trajectories to split into ρ groups.

Because we consider the natural frequencies of each oscillator are the same and by the following result, the global existence of all trajectories of Eq. (3) is ensured and moreover they will converge to fixed points.

Theorem 2. *The GKMF is a gradient system.*

As we have already mentioned, one of the properties of the GKMF is that the oscillators are naturally split into ρ groups (the order of the Fourier term).

Theorem 3. *Let $\mathbb{T}^n = \mathbb{S}^1 \times \cdots \times \mathbb{S}^1$ (n times) be the n-dimensional torus. Suppose that $K_F > K_N$ in Eq.* (3)*. Then the solutions of* (3) *converge to ρ stables subspaces of $\mathbb{T}^n$.*

Although the Fourier term forces the trajectories to split into ρ groups, it does not mean that this splitting is the best one, i.e., it does not mean that we will find the real communities only because of this splitting. Indeed, what the Theorem 3 says is that if K_F is too strong then the Fourier term will dominate the whole dynamics of the model and the topological contribution to the model will be lost. Therefore, the Fourier term can sharpen the detection of the communities for small values of K_F but can also worsen it for large values of K_F. Because of that, we must set small values for K_F in our numerical integration, otherwise the clustering will make no sense for community detection.

3 The Algorithm

Consider $\mathcal{G}$ a simple, non-directed and connected graph with n vertices and $\mathcal{C}_k$ a sub-graph of $\mathcal{G}$. The *inner degree* $g_i^{\to \mathcal{C}_k}$ of a vertex $i \in \mathcal{C}_k$ is the number $g_i^{\to \mathcal{C}_k} = \sum_{j \in \mathcal{C}_k} A_{ij}$, i.e, the number of links from i that are in $\mathcal{C}_k$. We define the *local clustering density* as

$$\delta_{\mathcal{C}_k} = \frac{\sum_{i \in \mathcal{C}_k} g_i^{\to \mathcal{C}_k} / g_i}{|\mathcal{C}_k|} \quad (4)$$

where g_i is the degree of the vertex i and $|\mathcal{C}_k|$ is the number of vertices inside $\mathcal{C}_k$.

If, in a process on community detection, ρ communities were identified then we define the *total clustering density* as

$$\delta(\rho) = \sum_{k=1}^{\rho} \delta_{\mathcal{C}_k}. \quad (5)$$

Accordingly to our given definition of community (Definition 1), a satisfactory community identification process must result in $\delta_{\mathcal{C}_k} > 0.5$ for all $k = 1, \cdots, \rho$ and as a consequence $\delta(\rho) > 0.5$.

As we said before, $\delta(2)$ is, probably, the larger total clustering density that we can find for almost every complex network regardless the number of communities. Therefore, it must be necessary to run the following algorithm for different values of ρ and an ultimate of a specialist must be necessary.

Algorithm to detect communities in complex networks via dynamics:

1 Evolve the phases of Eq. (3) choosing parameters accordingly to Theorem 1 and K_F preferentially $K_F < |K_N|$ and start setting $\rho = 2$.
2 In $\mathbb{S}^1$, split the phases into ρ groups.
3 Compute the local clustering density $\delta_{\mathcal{C}_k}$ for each $k = 1, \cdots, \rho$ and the total clustering density $\delta(\rho)$.
4 If each $\delta_{\mathcal{C}_k} > 0.5$ and, by consequence, $\delta(\rho) > 0.5$, the identified communities can be satisfactory. Repeat the above procedure changing the value of ρ to $\rho = 3, \cdots$ until having $\delta(\rho) \leq 0.5$. The correct community detection may have be decided by an expert.

4 Illustrations

The Zachary's karate club network is a social network with 34 members studied by Zachary [Zachary, 1977].

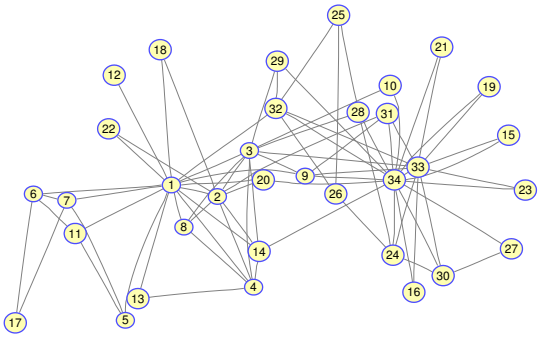

Figure 2. Zachary's karate club network.

First of all, we perform a comparison between the two presented dynamic models, namely, GKM and GKMF in order to provide numerical support for the use of GKMF. The Figure 3 depict it.

Lets then apply the algorithm presented in order to detect the communities in the Zachary's karate club network. The Table 1 present the results.

ρ	$\delta(\rho)$
2	0.902081
3	0.795850
4	0.448269

Table 1. Total clustering density for the Zachary's karate club network.

As we can see, both $\delta(2)$ and $\delta(3)$ are larger than 0.5 which lead us to conclude that this network can be well grouped into 2 or 3 communities. Indeed, some authors divided the karate club network into three communities [Shen et al, 2010]. Therefore, the algorithm has decided that this network can have at most 3 communities, but the final decision can be made by an expert

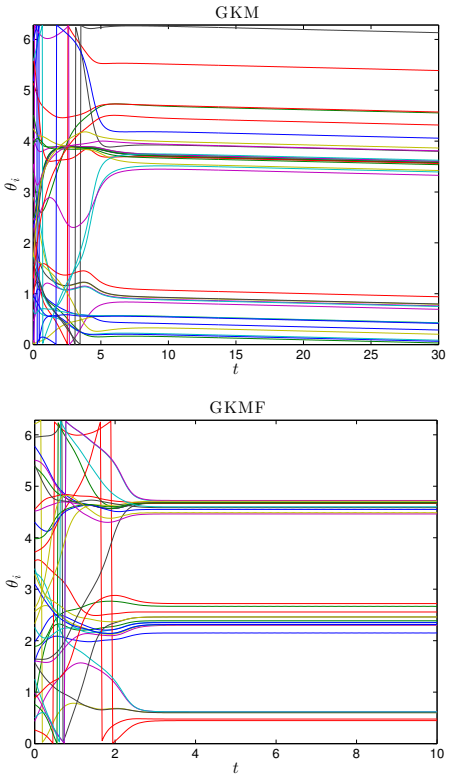

Figure 3. Comparative between the time evolution for GKM (top) and GKMF (bottom), with Fourier term of order 3, for the Zachary's karate club network. The parameters values used were $K_P = 34$, $K_N = -5$ and $K_F = 1$.

and it is known that in the social context this network has only 2 communities [Zachary, 1977]. The third identified community if formed by the set of vertices $\{5, 6, 7, 11, 17\}$ that arises from a larger community. The Figure 4 brings the Zachary's karate club network with the identified communities colored.

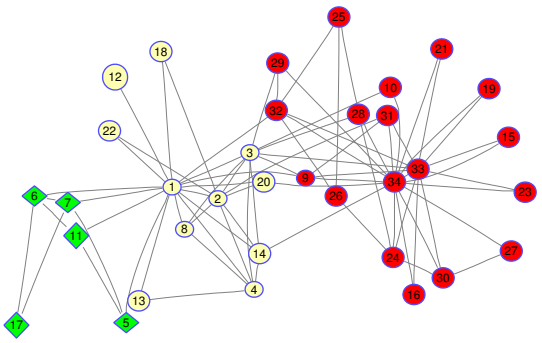

Figure 4. Zachary's karate club network with at most three communities identified.

5 Conclusions

Based on the GKM one have built a new model, namely, GKMF that have showed to be a little better

on community detection context rather than the GKM. Some results on this new model were presented and confirmed by numerical results. The Fourier term on GKMF can be a double-edged sword in the sense that its set up a natural clustering of the trajectories accordingly the order of this term, but if it is too strong the network structure contribution for the model can be lost. Numerical simulations have shown that for the most of network tested we should use $K_F < |K_N|$, but this value may be exceeded depending on the network size.

Acknowledgements

The authors would like to thank CNPq and FAPESP (project 2011/50151-0) for the financial support.

References

A. Arenas, A. Diaz-Guilera, and C. J. Perez-Vicente, *Synchronization Reveals Topological Scales in Complex Networks*, Phys. Rev. Lett. 96, 114102 (2006).

E. Oh, K. Rho, H. Hong, and B. Kahng, *Modular synchronization in complex networks*, Phys. Rev. E 72, 047101 (2005).

Fan Chung and Linyuan Lu. *Complex Graphs and Networks*. American Mathematical Society, 2006.

H. W. Shen, X. Q. Cheng, and B. X. Fang, *Covariance, correlation matrix, and the multiscale community structure of networks*, Phys. Rev. E 82, 016114 (2010).

J. Wu, L. Jiao, C. Jin, F. Liu, M. Gong, R. Shang and W. Chen, *Overlapping community detection via network dynamics*, Phys. Rev. E 85 (2012) 016115.

Jianshe Wu and Yang Jiao, *Clustering dynamics of complex discrete-time networks and its application in community detection*, Chaos, Vol. 24, 033104, 2014.

M. E. J. Newman, *Communities, modules and large-scale structure in networks*, Nature Physics 8, 2531 (2012).

Morris W. Hirsch, Stephen Smale and Robert L. Devaney. *Differential equations, dynamical systems, and an introduction to chaos*, Elsevier, 2004.

Santo Fortunato. *Community detection in graphs*. Physics Reports 486 (2010) 75-174.

S. Boccaletti. *Complex networks: Structure and dynamics*. Physics Reports 424 (2006) 175-308.

Schaeffer, Satu E., *Graph clustering*, Computer Science Review, V. 1, 27–64, 2007.

T. Nishikawa, F.C. Hoppensteadt and Y.C. Lai, *Oscillatory associative memory network with perfect retrieval*, Physica D 197 (2004) 134-148.

T. Zhou, M. Zhao, G. R. Chen, G. Yan, and B. H. Wang, *Phase synchronization on scale-free networks with community structure*, Phys. Lett. A 368, 431 (2007).

W. W. Zachary, *An information flow model for conflict and fission in small groups*, Journal of Anthropological Research 33, 452-473 (1977).

Y. Moreno, M. Vazquez-Prada, and A. F. Pacheco, *Fitness for synchronization of network motifs*, Physica A 343, 279 (2004).