Rubrique

Social Network Analysis

Novel method to find directed community structures based on triads cardinality

Gamgne Domgue Félicité* — Tsopze Norbert* — René Ndoundam*

* Computer Science Department - University of Yaounde I
BP 812 Yaounde - Cameroon
felice.gamgne@gmail.com, tsopze@uy1.uninet.cm, ndoundam@gmail.com

RÉSUMÉ. La détection des communautés est davantage un challenge dans les l'analyse des réseaux orientés. Plusieurs algorithmes de détection de communautés ont été developpés et considèrent la relation entre les nœuds comme symmétrique, car ils ignorent l'orientation des liens, ce qui biaise les résultats en produisant des communnautés aléatoires. Ce document propose un algorithme plus eff cace, TRICA, basé sur l'extraction des kernels qui sont des ensembles de nœuds inf uents dans le réseau. Cette approche découvre des communautés plus signif catives avec une complexité temporelle meilleure que celles produites par certains algorithmes de détection de communautés de l'état de l'art.

ABSTRACT. Community structure extraction is once more a major issue in Social network analysis. A plethora of relevant community detection methods have been implemented for directed graphs. Most of them consider the relationship between nodes as symmetric by ignoring links directionality during their clustering step, this leading to random results. This paper propose TRICA, an eff cient clustering method based on kernels which are inf uencial nodes, that takes into account the cardinality of triads containing those inf uencial nodes. To validate our approach, we conduct experiments on some networks which show that TRICA has better performance over some of the other state-of-the-art methods and uncovers expected communities.

MOTS-CLÉS : Réseaux orientés, détection des communautés kernel, Triade

KEYWORDS : Directed graphs, Community kernel detection, Triad.

1. Introduction

Community detection in directed networks appears as one of dominant research works in network analysis. The top meaning of community is a set of nodes that are densely connected with each other while sparsely connected with other nodes in the network [1]. This definition is interesting for undirected graphs ; like this many community detection algorithms implemented for directed networks simply ignore the directionality during the clustering step while other technics transform the directed graph into an undirected weighted one, either unipartite or bipartite, and then algorithms for undirected graph clustering problem can be applied to them.

These simplistic technics are not satisfactory because the underlying semantic is not retained. For example, in a food web network, according to them, the community structure will be corporated of predator species with their prays. This reflexion is not quite right. To make up for that idea, a generic definition of community detection consists of clustering nodes with homogeneous semantic characteristics(nodes centred around a set of objects owning the same interest). Our approach is based on extending the idea that within "good"communities, there are influencial nodes [6], ***kernels***, that centralize information, so that it will easily be attainable. Influancial nodes are crossed by a maximal number of triads in a community. A triad is a set of 3 nodes whose at least 2 are the *in-neighbor nodes* (target vertices) of the 3^{rd} vertex, or according to the triadic closure. Consequently, triads are the basis of many community structures [3]. Here we focus on the link orientation in triads. The specific contributions of our paper are :

– we mainly define a new concept named *kernel degree* to measure the strength of the pair of nodes and the similarity of vertices and give a new sense definition to kernel community based on the triadic closure.

– we develop a novel algorithm based on kernel degree to discover kernels and then communities from real social networks.

– We conduct to better quality improvement over the commmunity kernel detection algorithms.

The rest of paper is organized as follows. Section 2 is an introduction to related works. In Section 3, we formally define several concepts used into the proposed clustering method. In Section 4, we develop the algorithm. Section 5 is experiment study and Section 6 concludes this study.

2. Related works

Most approaches focused on symmetric models which lose the semantics of link directions, a key factor that distinguishes directed networks from undirected networks. For detecting communities in directed networks [2], some studies propose a simple scheme that converts a directed graph into undirected one, this enabling to utilize the richness and complexity of existing methods to find communities in undirected graphs, thus, to mesure cluster strength, they use an objective function, *the modularity*. Yet, this mesure has a limit resolution [1]. More recently, various probabilistic models have been proposed for community detection [7]. Among them, stochastic block models are probably the most successful ones in terms of capturing meaningful communities, producing good performance, and offering probabilistic interpretations. However, its complexity is enough because in pratice, if the number of iterations goes beyond 20, the method discontinue

and results become insignificant. To make up for this complexity, some authors define "kernels"like described below.

A *kernel* is considered as a set of influencial nodes inside a group. It seems to be information centralizing nodes. Some methods explored the problem of detecting community kernels, in order to either reduce the number of iterations, and consequently the time-complexity of algorithms defined for complex social networks or uncover the hidden community structure in large social networks. [4] identifies those influential members, *kernel* and detects the structure of community kernels and proposed efficient algorithms for finding community kernels. Through these algorithms, there is a random choice of the initial vertex, and the size of communities is fixed, leading to an arbitrary result estimation. To keep going, [3] proved that triangles(short cycles) play an important role in the formation of complex networks, especially those with an underlying community structure [5] and converts directed graph into an undirected and weighted one. This transformation misses the semantic of links. We propose a method which extracts triads based on Social properties to characterize the structure of real-world large-scale networks.

3. Method formalization

We propose in this section the kernel community model and introduce several related concepts and necessary notations.

3.1. Kernel community model

In directed networks, the link direction gives a considerable semantic to the graph and to the information flow. On twitter network for example, the notion of authority is pointed up as illustrated in Fig 1.(a), because of the relationship between a set of authoritative or hub blogs (nodes u and v) and a set of non-popular one called followers (nodes x) as presented in Fig 1.(b) and Fig 1.(c).

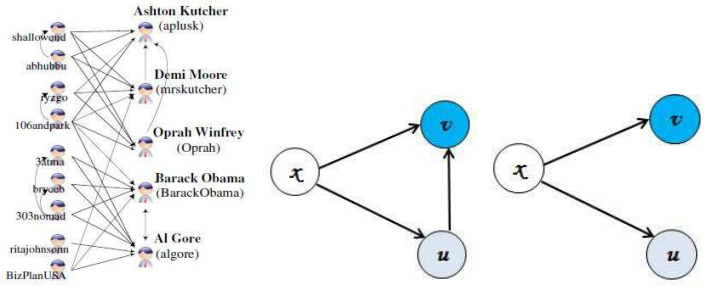
We integrated this concept of authority as one concept named **kernel degree**. Fig 1.(a) is a visualization of an extract from a twitter network. Kernel communities consist of nodes owning the same "in-neighbourhood "which corresponds to nodes that have more connections to the kernel (and not from the kernel) than a vertex outside the kernel. We consider only ingoing edges to the kernel vertices to express the strength these nodes get in some kind of network treated in this paper ; in a twitter network for example, hub blogs are viewed by many others followers and not the opposite ; in a citation network for example, authoritative authors like pioneers in a research area are more quoted by the others junior researchers. On the beginning, the kernel consists of two vertices sharing the same properties, leading to the notion of "triad "which consists of the idea that two vertices of the kernel share the same friend, like defined in the following sub-section.

3.2. Basic terminology and concepts

Given a directed graph $G(V, E)$ with $n = |V|$ vertexes and $m = |E|$ edges. Let Γ_u be the neighborhood vertices set of vertex u. We now give some following useful definitions :

Definition 1 (*Triad weight*). Let the identifier of vertex x in G be j. The triad weight of any edge (u, v) in graph G can be represented as Δ. We can use TW_{uv} to represent the number of triads (triad cardinality) crossing u and v according to the scheme presented in the Fig 1.(b) and Fig 1.(c).

$TW_{uv} = \frac{|\Delta_{uv}|}{|\Delta_j|}$.



(a) Illustration of Twitter Network (b) Closed triad (c) Opened triad

Figure 1. *Basic structures of our kernel community model.*

Definition 2 (*Neighborhood overlap*). Given two vertices u and v, let Γ_u be the set of vertices that are the neighborhood of vertex u, let Γv be the set of vertices that are the neighborhood of vertex v. Let NO_{uv} be the neighborhood overlap of u and v. $NO_{uv} = \frac{|\Gamma_v \cap \Gamma_u|}{|\Gamma_v \cup \Gamma_u| - 2}$ if there is an edge between u and v and 0 otherwise.

Definition 3 (*The kernel degree*). The Kernel degree of a pair of vertex u and v is : $K_{uv} = TW_{uv} * NO_{uv}$. K_{uv} can measure the strength of the pair (u, v) and the similarity of nodes.

Definition 4 (*New sense Kernel Community*). A new definition of the kernel community in the sense of this paper is a set of vertices with the same neighborhood such as these neighbors expand inward to the kernel, according the kernel degree K_{uv} gradually until its minimum.

Definition 5 (*Triadic Closure*). If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future.

The algorithm is structured into two steps : detecting kernel communities and then migrating the others vertexes to the kernel to whom they are more connected to.

4. Our Method for extracting communities

The new algorithm is structured in two steps : identifying kernels, then migrating the other vertices to the kernel as described in the following subsections. The algorithm for extracting Kernel communities, TRICA (Triads Cardinality Algorithm) we propose here makes use of a new concept *Kernel degree*, that measures the strength of a kernel gradually until it decreases. This concept is based on the triadic closure for emphasis the semantic proximity that links community members conducting to efficient propagation of information over the network. We focus on triads cardinality that is the number of neighboors two nodes own.

Data set	Vertices	Edges	Types
Extract from Twitter Network	14	31	Directed
American Football Network	115	613	Undirected
Celegansneural	297	2359	Directed

Tableau 1. *Data sets description*

4.1. TRICA algorithm

We assume that the network we want to analyze can be represented as a connected, directed, nonvalued graph G of $n = |N|$ nodes and $m = |E|$ edges. This step for identifying kernels is described in four sub-steps as follow :

1) Detect the *in-central* vertex v, which is the vertex with the maximal in-degree in the graph.

2) Determine the neighborhood overlap of each edge *(u,v)* through a variant of *Jaccard Index*[1] represented by NO_{uv} as defined in **Definition 2**

3) Store neighboorhood vertices u of v like $NO_{uv} > \varepsilon$

4) Compute K_{uv} through the *triad weight* TW_{uv} as described in **Definition 1**. This action is repeated to measure the strength of a kernel gradually until K_{uv} decreases.

These 4 substeps are repeated n/k times, k being the *in-degree* of vertex v.
The space complexity of TRICA is $O(n+m)$, and it runs in time more quickly than some of the state-of-the-art algorithms like shown in experiments.

The TRICA implementation for kernel communities is presented in Algorithm 1.

4.2. Deduction of global communities

After extracting kernels, it remains the other nodes which don't belong to the kernels ; they are called *non-kernels vertices*. The process of generating *global communities* (communities containing both kernels and non-kernels vertices) consists of migrating the other members (belonging to a set called "auxiliary communities") to the kernel whith whom they have a maximum number of connections, as described in Algorithm 2.

Algorithm 1 TRICA implementation for kernels extraction

Data: Directed graph $G = (N, E)$
Result: K Kernels
Initialisation : $K = \emptyset$;
repeat
  $k = d^{in}(v)/d^{in}(v) = max\{d^{in}(t), \forall t \in V\}$;
  Calculate NO_{uv} for each $(u, v) \in E$;
  $\Gamma_v[] \leftarrow \{t \in V/\exists t \in V, NO_{tv} > 0, 8\}$;$\Gamma_v[].sort$;$i \leftarrow 1$;
  $S \leftarrow \emptyset$;
  $j \leftarrow i$; $u \leftarrow \Gamma_v[j]$;$K^*_{uv} \leftarrow 0$;
  repeat
    Compute K_{uv} ;
    if $(K_{uv} > K^*_{uv})$ **then**
      $S \leftarrow S \cup u$;
    end if
    $u \leftarrow \Gamma_v[i++]$;
  until $(K_{uv} < K^*_{uv})$;
  $K \leftarrow K \cup S$;
until $(|V|/k)$
Return K ;

5. Experiments

To study the effectiveness and accuracy of TRICA, we compare it with following comparative methods :

– NEWMAN : Method for finding community structure in directed networks using the betweenness based on modularity [6].

Algorithms	Extract from Twitter		American Football		Celegansneural	
	%Δ	Comm Numb	%Δ	Comm Numb	%Δ	Comm Numb
Newmann	98%	2	39%	10	28%	194
Louvain	98%	2	63%	9	35%	5
Weba	98%	2	-	8	-	-
Triad Cardinality	**98%**	2	**70%**	12	**64%**	21

Tableau 2. *Community detection performance on the triad cardinality rate where the best rate are in bold.*

– LOUVAIN : Community detection algorithm based on modularity ; (we use Gephi tool for visualizing LOUVAIN results).

– WEBA [4] :Algorithm for community kernel detection in large social networks.

Algorithm 2 Algorithm implementation for non-kernels vertices migration

Data: Communities Kernels $K = \{K_1, K_2, ..., K_t\}$
Result: Global Communities $G_K = \{G_{K_1}, G_{K_2}, ..., G_{K_t}\}$
Let N be set of auxiliary communities ; $N = \{N_{K_1}, N_{K_2}, ..., NG_{K_t}\}$;
$\forall i \in \{1, ..., t\}, G_{K_i} = \emptyset$;
repeat
  $\forall i \in \{1, ..., t\}, G_{K_i} = K_i \cup N_{K_i}$;
  For $i \leftarrow 1$ **to** t **do**
    $S \leftarrow \{v \notin \cup G_{K_i} / \forall j \in \{1, ..., t\}\}$,
    $|E(v, G_{K_i})| \geq |E(v, G_{K_j})| > 0$;
    $N_{K_i} \leftarrow N_{K_i} \cup S$;
    $G_{K_i} \leftarrow K_i \cup N_{K_i}$;
  End For
until (No more vertices can be added)
Return G_K ;

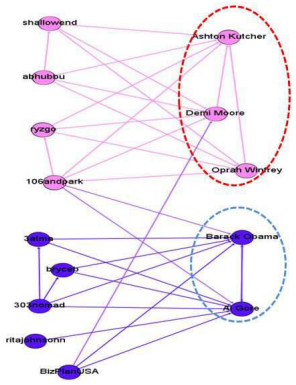
Our method is evaluated on directed and undirected networks. We use two levels of evaluation : The first is based on the time complexity, and the second on the **triad cardinality rate in communities**, that is the percentage of communities in the partition with highest triad cardinality rate. We use the function TCR defined as following, to evaluate our method :

$TCR = \frac{\Sigma_i |\Delta_i|}{|\Delta|}$

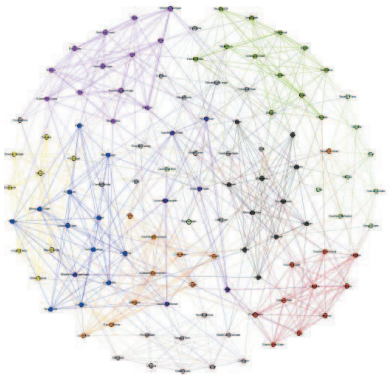
where i is one community and $|\Delta|$ the number of triads.

When we apply TRICA on the data sets described in the table 1, results in Fig 2 are following : The Fig 2.(a) illustrates the 2 expected communities of the Extract from Twitter Network, for all of the methods compared, with a triad cardinality rate in communities of 98% with kernels and followers [6]. But TRICA CPU time is better than other methods CPU time, as shown if 2.(c) The table 2 summarizes the comparison with some state-of-the-art methods. It shows that Triad Cardinality algorithm provides the highest triad cardinality rate in communities. As far as the Football network is concerned, Triads cardinality algorithm can divide the network into 12 communities exactly as shown in Fig 2.(b). In this result, 8 communities are completely consistent, this revealed by the triad cardinality rate of 70%. Meanwhile Newmann algorithm can divide it into 10 communities and LOUVAIN into 9. This number of communities does not reflect the real structure of the American College Football network. On the other hand, the result for applying

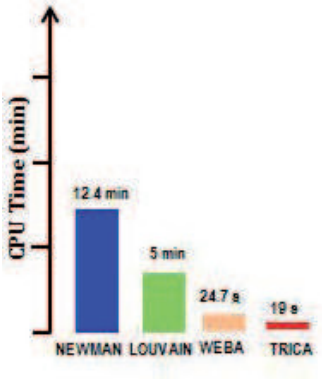
TRICA to Celegansneural network shown in Table2 presents that TRICA detects 21 communities, while LOUVAIN detects 5 and NEWMAN 194. But the triad cardinality rate is the best, 64%, certifying that our method uncovers a better structure of social networks.



(a) Extract of Twitter Network



(b) American College Football network.



(c) Efficiency comparison of TRICA and others algorithms on Twitter Network.

Figure 2. *Results of applying TRICA to data sets.*

6. Conclusion

In this paper, we focus on the problem of kernel community detection in directed graphs, kernels being the key tool for understanding the role of networks and its structure. We mainly interested on extracting kernels which are influential nodes on the network. Our kernel community model define triads according to some social properties to characterize the structure of real-world large-scale network, and we develop a novel method based on the proposed new concept, the *kernel degree* which defines the strength of kernel community. Experiments proved that TRICA detects efficiently expected communities and achieves 20 % performance improvement over some other state-of-the-art algorithms, but it only works for unweighted graphs. Our next work is to optimize Triad cardinality-

based property, and adjust it to suit for detecting kernel communities from large-scale directed and weighted networks.

7. Bibliographie

[1] S. FORTUNATO, « Community detection in graphs », *Physics Reports* 486(3) 75-174, 2010.

[2] F. D. MALLIAROS and M.VAZIRGIANNIS, « Clustering and community detection in directed networks : A survey. », *arXiv* 1308.0971, 2013.

[3] C.KLYMKO , D.F GLEICH and T.G KOLDA, « Using Triangles to Improve Community Detection in Directed Networks », *Conference Stanford University*.

[4] LIAORUO WANG , TIANCHENG LOU , JIE TANG and JOHN E. HOPCROFT, « Detecting Community Kernels in Large Social Networks ».

[5] A. PRAT-PÉREZ , D. DOMINGUEZ-SAL , J. M. BRUNAT and J. L. LARRIBA-PEY, « Shaping communities out of triangles. », *In CIKM 12* n° 1677-1681, 2012.

[6] FÉLICITÉ GAMGNE and NORBERT TSOPZE, « Communautés et rôles dans les réseaux sociaux », *in : CARI '14 : Proceedings of the 12th African Conference on Research in Computer science and Applied Mathematics* n° 157 - 164, 2014.

[7] TIANBAO YANG , YUN CHI , SHENGHUO ZHU and YIHONG GONG and RONG JIN,« Directed network community detection : A popularity and productivity link model. », *In SIAM Data Mining'10* n° 2010.