GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

COMMUNITY DETECTION IN MOBILE SOCIAL NETWORKS

L.Boujlaleb[*1], D.Mammass[2], A.Idarrou[2] and S.Idrissa[3]
[*1,2]University Ibn Zohr, Agadir
[3]University Cheikh Anta Diop, Senegal

ABSTRACT

Given the growth of the availability of location-acquisition technologies, we see nowadays new types of social networks, called mobile or location-based social networks. These types of networks can be represented by attributed graphs whose nodes are described by the transient attributes' data that are often geographic or contextual and edges represent the relationships between nodes. The analysis of these networks is based on the study of the structure of communities. However, most of the community detection methods mainly focus on the structural information of the network. This paper, presents a state of art of dynamic community detection methods in social networks that integrate both the structure of the network along with attributes describing the node. At first, we present our approach adapted to mobile network based on the spatio-temporal character. We then show that the use of attributes describing the nodes enhances the relevance of the results and provides information closer to reality.

Keywords- *Location-based social networks, dynamic detection of communities, spatial-temporal attributes, attributed graph.*

I. INTRODUCTION

Social networks are used as facilitators and sometimes precursor means of interaction and exchange between geographically separated people. Indeed, social networks offer opportunities for communication, data and knowledge sharing, identification of new employees, etc. To take advantage of social network's uses, these later should be analyzed to better identify their components, their characteristics, their evolution, groups or communities that emerge among them, etc.

With the proliferation and democratization of smart phones with GPS (Global Positioning System) and location-based services, we see nowadays new types of social networks, called mobile or location-based social networks.

With mobile social networks, data can be of two types: permanent data (who is friend with whom, affiliations, web links, etc.) and transitional data that are often geographic or contextual which come from discrete events such as publications, recommendations of favorite places in a geographic area, ad hoc communities, etc.

The analysis of a social network is often done on data stored using a graph structure called attributed graph where the nodes are the entities that are associated to a number of attributes that describe them and the edges represent the interaction between them.

To understand and analyze these systems, we study their communities' structure. The graphs consist of a set of nodes strongly interlinked and weakly bound to the rest of the network.

Most social network analysis methods use only permanent data (structural information), while the use of transitional data (relational information) enhances the relevance of the results and provides information more close to reality.

This approach is particularly explained by the phenomenon of homophily which means that individuals who have similar attributes tend to bind among each other. The homophily behavior can be observed in many complex networks such as the social network, the network of citation and other types of networks [12].

In this paper, we present a state of art of community detection methods in attributed graphs. We then propose a hybrid detection method suitable for mobile social network whose nodes are described by spatial and temporal data.

We begin with a brief description of the traditional community detection methods. We then describe methods for the detection of communities in attributed graphs. For this, we adopt a classification to categorize different communities' detection methods that take into account the network structure and the properties that describe the vertices; the first category includes methods for detecting communities that rely on the extension of the method of [6]. The second category is the integration of attribute values in the original graph. The third category is based on the statistical model.

Section 2 is devoted to the state of art. Section 3 introduces the preliminary concepts and formulates the problem of classification of attributed graphs suited for location-based social networks. We end with a conclusion and perspectives.

II. STATE OF ART

2.1 Method based on extending method of [6]

To extend the modularity of Newman, in [3], a method is proposed in order to add a term to measure the similarity based on the attributes of two nodes. Structural and attributes information are considered simultaneously during the partitioning process. In [1], the authors proposed two algorithms SAC1 and SAC2: algorithm SAC1 checks repeatedly all nodes leading to complexity O (n2). In order to reduce the computational cost, they have proposed an alternative approach based on node's nearest neighbors.
[2] Also worked on a community detection method derived from the method of [6] and based on the optimization of two criteria: the modularity for structural relations between nodes and the entropy for attributes. The entropy of the partition is optimized following the Monte-Carlo approach [2]. The algorithm follows the principles of [6] but by optimizing successively the entropy and the modularity in order to minimize the first and maximize the second.

[8] Proposed a new method, ToTeM which optimizes a global criterion composed of two functions: the modularity for structure and inter-class inertia for attributes. This method resembles to the SAC1 method of [1] in dealing with both relations and attributes during the partitioning except that SAC1 optimizes the similarity between pairs of vertices belonging to the same partition while To TeM optimizes the inter classes inertia measure.

2.2 Methods based on the integration of attribute values in the original graph

[3] Proposed SA-Cluster, an algorithm for detecting communities which includes the structure and attributes describing the nodes. A set of vertices-attribute and edges-attribute are added to the original graph which leads to the creation of a new graph called "Attribute augmented graph". The edges-attributes are added between each vertex-attribute and the original vertices of the graph and take the value of vertex-attribute. With such augmentation, the attribute similarity is computed by the proximity of vertices in the graph: two vertices sharing attribute value are connected by vertex-attribute. A partitioning is performed on the attribute augmented graph.
To improve the efficiency and Scalability of SA-Cluster method, [4] proposed an incremental version of Inc-Cluster which updates progressively the random walk distances.

In [10], the authors propose a method where the similarity between each pair of vertices is computed based on attributes and relational proximity of vertices. This similarity measure is used as link weight between each pair of vertices. Then, any graph portioning method may be applied to the weighted graph [10].

In [11], an approach provides a framework for recommendation links for forecasting links. This method consists of adding new vertices representing the attribute values under the name (social- attribute network, SAN).

2.3 Methods based on statistical model

In [9], the method proposed is based on laws to model the attributes distribution and relationships in networks by assessing the probability that two nodes will be classified.

2.4 State of art methods analysis

Through the study of the three categories of community detection methods that consider both the relationship between nodes and their attributes, we have identified the following conclusions:
Methods extending the method [6] are based on the optimization of local criteria or leading to conflicting choices. For example, [2] propose an algorithm incorporating the principles of [6] but successively optimizing entropy and modularity in order to minimize the first and maximize the second. Conflicting changes may be caused due to these two successive optimizations.

Instead of successively optimizing the two criteria; ToTeM optimizes a criterion taking into consideration the overall quality of the partition related to attributes and relationships. This method resembles to SAC1 method [1], however, ToTeM improves a measure of inter-class inertia and SAC1 examines the similarity between pairs of vertices of the same class.

The methods of the second category using a graph summarizing the two types of information are characterized by inadequate to discrete data because it is necessary to introduce a lot of vertices. Besides this, the orderliness of attribute values is not considered. These methods ignore certain distances between elements during the classification.

The last category is based on the statistical methods that are costly in computation time or making assumptions that are not suitable for all mobile networks.

All these methods have limitations for mobile social networks because they do not manipulate the spatial and temporal attributes that describe the nodes.

Our approach is inspired from the method of [6] because we find that this method is characterized by its quick execution, the ability to be used in very large networks and the number of clusters to form is not required prior the execution because it is not supervised. Finally, the difficulty of optimizing the modularity is proven NP-complete; the method of [6] is a heuristic optimizing the criterion which gives a suitable solution in a reasonable time [6].

III. PROBLEM STATEMENTS

3.1 Attributed graph:

An attributed graph is denoted G = (V, E, X)

- V is a set of vertices,
- E is a set of edges,
- X= $\{a_1,\ldots,a_m\}$ is a set of attributes associated with vertices in V to describe the node properties.

Each vertex vi • V is associated to a vector of attributes $[a_1(v_i),\ldots, a_m(v_i)]$ where $a_j(v_i)$ is the attribute value of the vertex v_i for the attribute a_j.

The goal is to find communities in the attributed graph which partition the graph into disjoint groups in r distinct $P=\{C_1,\ldots,C_r\}$ where r is a priori unknown and the partitions do not overlap and each class has at least one element in Gi=(Vi,Ei,X) where

- $\bigcup_{k=1}^{r} C_k = V$
- $C_k C_l =$
- $C_k, k\{1,\ldots,r\}$

We propose a new approach that simultaneously considers the topological structure of the graph and the attributes describing the vertices during the partitioning. This method is an extension of the method of [6], which is based on criteria of modularity. Our algorithm is suitable for mobile social networks; it uses two notions simultaneously: the modularity which classifies nodes based on relationships and the « hierarchical density classification based on spatial and temporal dimension » method which considers the geographical and temporal attributes of nodes. We want that the impact of the two sources of information (structural and relational) to be balanced during the portioning process. In other words, our method must produce clusters whose nodes are strongly interrelated and less associated with nodes of other clusters and the nodes' attributes of the same cluster to be similar.

3.2 Classification based on relationships: the modularity of Newman Q_{Newman}

The modularity is a metric used to measure the quality of a division into communities.
It is the difference between the proportion of edges inside a community and the proportion of links in an expected community in a random graph of the same degree of distribution.
Given a graph with n nodes and m represents the number of edges, Gi,j represents the link (i,j), di : degree of node i. if the graph is partitioned into K partitions, the modularity of [5] is computed as follows (1) :

$$Q_{Newman} = \sum_{l=1}^{K} \sum_{i \in C_l, j \in C_l} S(i,j) \qquad (1)$$

Where the link strength S (i,j) between two nodes i and j is computed by comparing the actual interaction of the network Gij with the expected number of connections (di . dj)/2m

$$S(i,j) = \frac{1}{2m} . (G_{i,j} - \frac{d_i . d_j}{2m}) \qquad (2)$$

3.3 Classification based on spatial temporal attributes: modularity of spatial temporal attributes" Q_{AttrST}

The modularity of Newman does not include the similarity between nodes' attributes. Our contribution is to add a term that takes into account the spatial temporal character during the partitioning. To do this, we will define a "modularity of spatial temporal attributes" of a partition as follows (3):

$$Q_{AttrST} = \sum_C \sum_{i,j \in C} simA(i,j) \qquad (3)$$

Where *simA* is the attribute similarity function.

To compute the similarity function *simA* between nodes considering the geographic and temporal attributes, we use the « hierarchical density classification based on the spatial and temporal dimension » method [7].

Step1 : Detection of spatial temporal partitions from the check-ins of nodes at different levels of spatial temporal hierarchy.

This step is based on the classification algorithm ST-DBSCAN (Space and temporal Density-based spatial clustering of applications with noise). This algorithm uses three input variables: spatial dimension (Eps Space), temporal dimension (Eps time) and the minimum number of POI (point of interest).

The size of the partition changes gradually from large size to small one and allow having similar partitions at different levels of spatial temporal hierarchy as explained in table 1.

Table I. Partitions at different levels

Level	Eps_space/km	Eps_time/hour	MinPts
4	4	3	40
3	8	6	160
2	16	9	640
1	32	12	2560

Step2: For each spatial temporal level, the cosine similarity between nodes is computed using the space vector model with vectors composed of the number of visits for each node in different spatial temporal partitions.

In other words, a node-check-in matrix represents spatial and temporal dimension in a region of a classification considering all nodes in a given period is defined as:

$$V_{l(m\,x\,n)} = \begin{bmatrix} V_{1,1} & \cdots & V_{1,n-1} V_{1,n} \\ \vdots & & \vdots \quad \vdots \\ V_{m,1} & \cdots & V_{m,n-1} V_{m,n} \end{bmatrix}$$

where m is the total number of visiting nodes, n is the number of partitions discovered by ST-DBSCAN (Eps_space, Eps_time, MinPts), V_{ij} is the number of check-ins of node i in the partition region h, and l is the level detail in classification hierarchy.
The location is considered as a vector in n-dimensional space. The cosine angle between two vectors is used in order to measure the similarity between nodes. Suppose that node i and node j are represented by vector Ui and Uj in n-dimension check-in space.
The similarity between nodes i and j is defined as follows:

$$simA(i,j) = \cos(U_i, U_j) = \frac{U_i \cdot U_j}{||U_i|| \, ||U_j||} \qquad (4)$$

Step3: the global similarity of partition's nodes in each spatial-temporal hierarchy levels is computed as follows:

$$\text{simA}_{\text{overall}} = \sum_{i=1}^{N} \mu \text{sim}_i \qquad (5)$$

$$\mu = \beta_i \Big/ \left(\sum_{i=1}^{N} \beta_i\right) \qquad (6)$$

3.4 Composite modularity Q

Then, we will introduce a composite modularity Q (7) as weighted combination of the« Newman modularity» Q_{Newman} (1) and the « modularity of spatial temporal attributes » Q_{AttrST} (3) :

$$Q = \sum_C \sum_{i,j \in C} \alpha . S(i,j) + (1-\alpha) . simA(i,j) \qquad (7)$$

α *is the weighting factor*, $0 \leq \alpha \leq 1$

Next, an approximate optimization of composite modularity Q should be found following directly the approach inspired by algorithm of [6].
[6] Uses a greedy method for optimizing modularity. The algorithm starts with each node belonging to its own community. It chooses randomly a node and attempt to move it to the community of its neighbors. If a positive gain is obtained, the node is placed in the community with the maximum gain. If not, it remains in its original community. This step is applied until there is no more improvement.

Algorithm:

Inputs: Attribute graph G =(V,E,X) and similarity matrix
Outputs: set of communities
Step1 : Initialize each node in a single cluster
repeat
  for all vertices i in V **do**
    for all j neighbors of i in V **do**
      Move i to j's community
      Compute the composite modularity
      gain Q
    End
    Choose j with maximum and positive
    Gain (if exists) and
    Move i to j's community If not i stays in
    its initial community
  End for
Until maximization of the modularity

IV. CONCLUSION & PERSPECTIVES

In this paper, we have studied community detection methods in social networks which are the most related to our topic. At first, we proposed a hybrid community detection method suitable for mobile network whose nodes are described by spatial and temporal data. As perspectives, we will conduct experiments on several real social networks, and we will apply different methods to compare the number of communities, communities' size and modularity for structure and for attributes.

REFERENCES

1. *The Anh Dang and Emmanuel Viennet. Community Detection based on Structural and Attribute Similarities. International Conference on Digital Society (ICDS), pp. 7-14, 2012.*
2. *Juan David Crus-Gomez, Cécile Bothorel, and François Poulet. Entropy based community detection in augmented social networks. CASoN, pp. 163-168, 2011.*
3. *Yang Zhou, Hong Cheng, and Jeffrey Xu Yu. Graph clustering based on structural/attribute similarities, VLDB, pp.718-729, 2009.*
4. *Yang Zhou, Hong Cheng, and Jeffrey Xu Yu. Clustering Large Attributed Graphs: An efficient Incremental Approach, ICDM, pp.689-698, 2010.*
5. *Aaron Clauset, Mark Newman, and Cristopher Moore. Finding community structure in very large networks, Physical Review E, vol. 70, p. 066111, 2004.*
6. *Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment,vol. 2008, no.10, p.P10008(12pp), 2008.*
7. *Yuan Zhengwu, Jiang Yanli, Gidofalvi Gyozo. Geographical and temporal similarity measurement in location-based social networks. MobiGIS, 2013.*

8. *David Combe, Christine Largeron. ToTeM: une méthode de détection de communautés adaptées aux réseaux d'information. EGC, 2013.*
9. *Zhiqiang Xu, Yi Wang, Hong Cheng. A model-based approach to attributed graph clustering. SIGMOD, 2012.*
10. *Karsten Steinhaeuser, Nitesh Chawla. Community Detection in a Large Real-World Social Network. Social Computing, Behavioral Modeling, and Prediction, p. 168-175, Springer US, 2008.*
11. *Zhijun Yin, Manish Gupta, Tim Weninger, and Jiawei Han. LINKREC: a unified framework for link recommendation with user attributes and graph structure. Proceedings of the 19th international conference on World wide web- WWW'10, New York, USA, pp. 1211. ACM Press, 2010.*
12. *Aris Anagnostopoulos, Ravi Kumar, Mohammad Mahdian. Influence and correlation in social networks. Proc. 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.7-15, 2008.*