Community Detection from Location-Tagged Networks

Zhi Liu, Yan Huang
Department of Computer Science and Engineering
University of North Texas, Denton, Texas
zhiliu@my.unt.edu, huangyan@unt.edu

ABSTRACT

Many real world systems or web services can be represented as a network such as social networks and transportation networks. In the past decade, many algorithms have been developed to detect the communities in a network. However, the impact of locations on community has not been fully investigated by the research literature. In this paper, we propose a method to determine if a location-based community detection method is suitable for a given network and provide a new community detection algorithm that pushes the location information into the community detection. We test our proposed method on both synthetic data and real world network datasets. The results show that the communities detected by our method distribute in a smaller area compared with the traditional methods and have the similar or higher tightness on network connections.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

Keywords

Community Detection, Geo-tagged Network, Connection Locality

1. INTRODUCTION

Many real world systems or web services can be represented as a network such as social networks, the World Wide Web, and biological networks. Detecting communities from networks has received considerable attention and is the main focus of many research efforts in the past decade [1, 4, 2]. Generally, the goal of community detection is to find the subgraphs with tight internal connection based on node connections, labels of nodes, and the weights. However, the formation of many real world networks is greatly influenced by the geographic locations of the nodes which has not been fully investigated by the currently literature. We observe that the nodes in a tightly connected community tend to be more close to each other in space as well. Location can have different impact on social networks and the impact can be quantified and used in community detection. Introducing locations of nodes to community detection can improve the performance of detection on real world networks. In this paper, we propose community detection methods that take the locations of the nodes into consideration with the main goal of improving the quality of the detection results in terms of average internal degree, accuracy, and geographic span of detected communities.

We focus on finding communities with nodes distributing in a small range of area and at the same time, keeping the connection tightness of the nodes in the community. This paper makes the following contributions: 1) Developing the algorithms to detect communities with locality on large location-tagged networks; 2) Given a location-tagged network, we proposed a new measurement called Total Variation Difference to help determine if the network has a locality property and a location-based community detection method is suitable. 3) We propose optimization techniques and indexing method to allow the algorithm to scale well for large networks. It took around 30 seconds to detect communities from a real network of 20,000 nodes; 4) We test our proposed method on both synthetic data and real world network datasets. The results show that the communities detected by our method distribute in a smaller area compared with the traditional methods and have the similar or higher tightness on network connections. In the following sections, we will introduce some related works(section 2), our method(section 3), the experiment results(section 4), and the conclusion (section 5).

SIGSPATIAL '14, November 04 - 07 2014, Dallas/Fort Worth, TX, USA

http://dx.doi.org/10.1145/2666310.2666496

2. RELATED WORK

Community detection: In the past decade, many algorithms have been developed to detect communities in a network. For complete discussion of various algorithms, please refer to [2]. Aaron *et al.* provide a hierarchical clustering approach to detect communities using internal density in [1]. The internal density is the number of edges inside a community in a network. The basic idea is to increase the ratio of the edges in communities during the hierarchical clustering process using Equation 1:

$$Q = \frac{1}{2m}\sum_{vw}[A_{vw} - \frac{k_v k_w}{2m}]\delta(c_v, c_w) \quad (1)$$

where A_{vw} is the adjacency matrix of the network and k_v is the degree of node v. c_v represents the community of n-

ode v and $\delta(c_v, c_w)$ is 1 if $c_v = c_w$. m is the number of edges in the whole network G. Another popular algorithm [4] is based on iteratively removing "unimportant" edges. The basic assumption of this method is that communities are weakly connected by a few edges. In [3], the authors define the similarity between nodes using their degrees and the number of common neighborhood. The sum of the similarities of edges inside or outside a community was defined as internal or external similarity of a community. These works do not consider locations of nodes in a network.

In the last few years, some researchers have studied the geographic constraints on real world networks. In [5], the authors build a network based on the cell phone communication records. Then they study the relationship between distance and the call/text tie probability. In [6], the authors define the concepts of node locality and geographic clustering coefficient. Then they show the value distribution of these two coefficients with respect to the degree of nodes. Their study shows that people tend to build connections with other nearby users.

3. THE ALGORITHM

We denote the network as $G = (N, E, L)$, where N is the set of nodes, E is the edge set, and L is the location set of the nodes. To determine whether the locations of nodes will help in community detection, we will analyze the locality of the network first. Then we propose our locality-based method. We follow the hierarchical clustering framework combined with the location information. A good division of the network produces communities with higher ratio of internal edges and smaller geographical scope.

3.1 Network Locality

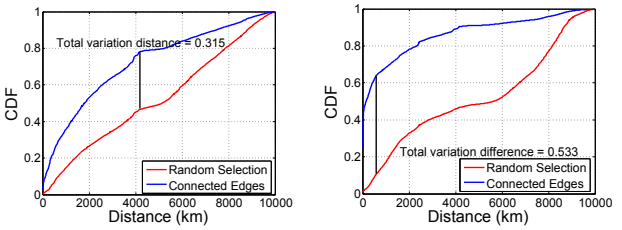
The formation of connections in many real world networks are influenced by the location of nodes in the network. However, some networks are more location influenced than others. So before we provide the location-based community detection algorithm, we need to analyze the influence of the location on networks to see the degree of influence. This will be helpful in determining if location based community detection is a suitable method. Here, we use network locality defined below to measure the relationship between location and connection in a network.

Definition 1 (Network Locality). *In a network G, we use two indexes to measure its locality: Total Variation Difference (TVD) and the Inflection Distance. Let $F(dis)$ be the cumulative distribution function (CDF) of distance between any two nodes in G and $F_c(dis)$ be the CDF of the distance between connected nodes in G, the total variation distance is defined as:*

$$TVD(F, F_c) = max(F_c(dis) - F(dis)) \quad (2)$$

and the Inflection distance is defined as the distance where $F_c(dis) - F(dis)$ achieves the maximum value.

We can see that a higher value of the total variation distance indicates the network is more geographically close because connected nodes in nearby locations have higher percentages. When the total variation difference is close to zero, the connection has little relationship with the locations of nodes. We analyze the network locality of two real datasets:



(a) Twitter: the total variation distance is 0.315 and the inflection distance is $4180km$

(b) Gowalla: the total variation distance is 0.533 and the inflection distance is $580km$

Figure 1: The cumulative distribution function of distance between every user pair/friend pair on Twitter and Gowalla.

Gowalla and Twitter. The Gowalla dataset is a 99,563 users' friendship network and the Twitter dataset contain 148,860 users. In Figure 1, we plot the cumulative distribution function of distance between every user pairs and friend pairs. In the Twitter dataset, the total variation distance is 0.315 and the inflection distance is $4,180km$. Compared with Twitter, the Gowalla network is more close geographically since it has a higher TVD, 0.533, and a smaller inflection distance $580km$. This phenomenon illustrates that users in Gowalla tend to build friend relations with others who are geographically close to them compared with Twitter. In other words, the locations of nodes have greater influence on the network structure in Gowalla. We suggest applying our method on the networks with the TVD larger than 0.25.

3.2 Connection Locality

To take location into account in community detection, first we define the concept of connection locality to qualify the graphic closeness between nodes.

Definition 2 (Connection Locality). *Let dis_{vw} be the geographic distance between nodes v and w. Let σ be the average distance between all user pairs. The connection locality can be defined as $L_{vw} = exp(-dis_{vw}/\sigma)$.*

And then we measure the geographic and network closeness of the communities using the following equation:

$$C_G = \frac{1}{\sum_{vw} A_{vw} L_{vw}} \sum_{vw} A_{vw} L_{vw} \delta(c_v, c_w) \quad (3)$$

We can see that this method is equivalent to assign each edge in network G with the locality as weight. Inspired by the method in [1], we introduce the expected value of C_G to avoid the situation that the largest value of C_G will be achieved when all the nodes belong to the same community. The expected value of C_G is obtained from a random connection network. The probability of an edge existing between a node pair is $k_v k_w / 2m$. Since we already know the locations of the nodes, the expect value for each edge is $l_{vw} k_v k_w / 2m$ and the expect value of C_G is the sum of the expect value of all the edges. Let $\omega = \sum_{vw} A_{vw} L_{vw}$, we define the modularity Q as:

$$Q = \frac{1}{\omega} \sum_{vw} [A_{vw} L_{vw} - \frac{k_v k_w}{2m} L_{vw}] \delta(c_v, c_w) \quad (4)$$

3.3 Node Similarity

To enhance the influence of network structure, here we define the node similarity between nodes pair by the common neighbors and their degrees:

DEFINITION 3 (NODE SIMILARITY). *Let Γ_v be the set of neighbors of vertex v. The similarity of two nodes is calculated by their common neighbors and their degrees as:* $S_{vw} = |\Gamma_v \cap \Gamma_w| / \sqrt{|\Gamma_v||\Gamma_w|}$

To apply the node similarity in our modularity, we also need to calculate the expect value under random connection network. Assuming node i has k_i neighbors, the probability of node i is connected to v (w) is $k_v k_i/2m$ ($k_w k_i/2m$). So the probability of node i connected to both v and w is $\frac{k_v k_i}{2m}\frac{k_w k_i}{2m}$. The expected value of S_{vw} is the sum of the probabilities of both v and w connected to any other node i:

$$S_{vw} = \frac{|\Gamma_v \cap \Gamma_w|}{\sqrt{|\Gamma_v||\Gamma_w|}} = \sqrt{k_v k_w} \sum_{i \neq v \& i \neq w} k_i^2/4m^2 \quad (5)$$

In practice, we use $\tau = \sum_i k_i^2/4m^2$ instead of $\sum_{i \neq v \& i \neq w} k_i^2/4m^2$ because they have similar value on larger networks.

We then revise ω as $\sum_{vw} A_{vw} S_{vw} L_{vw}$, and the new modularity Q^s is defined as:

$$Q^s = \frac{1}{2\omega} \sum_{vw} [A_{vw} S_{vw} L_{vw} - L(v,w) \frac{k_v k_w}{2m} \tau \sqrt{k_v k_w}] \delta(c_v, c_w) \quad (6)$$

In this paper, we only consider the node similarity between connected nodes for the following reasons: (1) Relation of 2-degree neighbors (the node pairs which are connected but share at least one common neighbors) introduce many new connections. The number of 2-degree neighbors is much more than directly connected neighbors and will significantly increase the computation complexity. (2) The influence of 2-degree neighbors is much smaller than directly connected ones. Based on our investigation, the average distance between 2-degree neighbors are three to times times longer than directly connected neighbors even when they have the same node similarity.

3.4 Optimization

In [1], the authors provide an efficient method to implement their model. They maintain and update a matrix ΔQ_{ij} which records the change of Q after combing the communities i and j. We can rewrite the modularity in Equation 6 into Equation 7. By analyzing the modularity, we can see that after we combine two communities i and j, the change of Q^s includes two parts: 1) the connections between these two communities will increase the value of Q^s (the first part in Equation 7) and 2) the value generated by node pairs from communities i and j (the second part).

$$\begin{aligned} Q = & \frac{1}{2\omega} \sum_{vw} A_{vw} S_{vw} L_{vw} \delta(c_v, c_w) \\ & - \frac{1}{2\omega} \sum_{vw} L(v,w) \frac{k_v k_w}{2m} \tau \sqrt{k_v k_w} \delta(c_v, c_w) \end{aligned} \quad (7)$$

The combination of two disconnected communities will not increase the value of Q, so we only keep the ΔQ_{ij} if there is at least one edge between them. At first, every node is a community and the ΔQ between each connected node pair is: $L_{ij}[\frac{S_{ij}}{2\omega} - \frac{\tau (k_i k_j)^{1.5}}{4\omega m}]$. After we combine communities i and j, we need to update all the communities k which are connected to i or j. We use (ij) to denote the community generated by combining i and j and use $\Delta Q_{k,(ij)}$ to denote the new ΔQ value between k and (ij). If the community k is connected to both i and j, we can get the new $\Delta Q_{k,(ij)}$ by $\Delta Q_{ik} + \Delta Q_{jk}$. If k is only connected to one of them, e.g. i, we do not have Q_{jk} since they are disconnected. So we need to calculate it as:

$$\Delta Q_{jk} = -\frac{1}{2\omega} \sum_{vw} \tau L(v,w) \frac{(k_v k_w)^{1.5}}{2m} \delta(c_v, j) \delta(c_w, k) \quad (8)$$

And then we can update the $\Delta Q_{k,(ij)}$ by:

$$\Delta Q_{k,(ij)} = \Delta Q_{ik} + \Delta Q_{jk} \quad (9)$$

We will stop the hierarchical clustering process when the modularity Q achieve its maximum value, which means that the largest ΔQ_{ij} is less than zero.

4. EXPERIMENT

In this section, we test our method on synthetic networks and two real world social network datasets described before: Twitter and Gowalla. We use three different measurements to evaluate the results: 1) Geographic Span: the average distance of the nodes in community c to the centroid $(\bar{x}, \bar{y})$ of all the nodes in this community; 2)Average Internal Degree: the internal degree of a node v means its neighbors in the same community. The average internal degree of a community c is the average value of the internal degrees of all the nodes in c. 3) Accuracy: Since we do not have a class label of the real world datasets, we only apply this on the synthetic networks. We implemented four community detection methods in our experiments: 1) Randomly select nodes as community (Random). 2) The method proposed in [1] (Clauset's Method). 3) The method discussed in section 3 using Equation 4 as the modularity Q (Connection Locality). 4) The method discussed in section 3 with Equation 6 as the modularity (Node Similarity).

Table 1: Accuracy of different community detection methods

Ω	Clauset's	Connection Locality	Node Similarity
1	16.24	16.63	18.38
3	16.48	22.82	24.63
5	17.72	22.40	28.77
10	22.16	25.14	26.60
30	32.84	19.76	24.42
$+\infty$	36.04	19.20	19.76

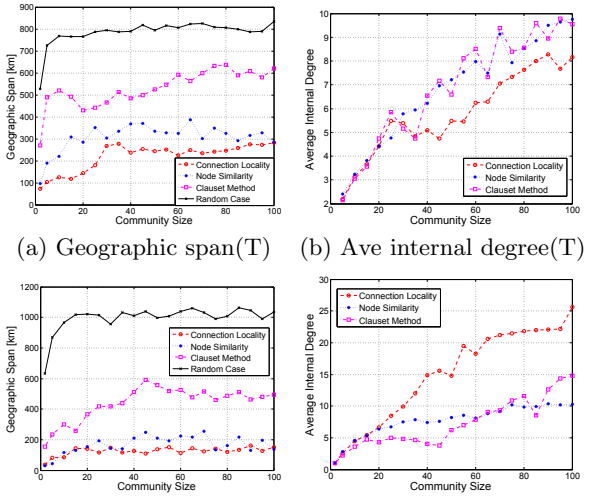
4.1 Tests on Synthetic Networks

First we test the methods on the generated networks because a synthetic datasets allow for better parameter control. We generate the networks on a 50×50 grid with 2,500 nodes in it. For each node, we randomly assign a community label to it and there are 10 different community labels. We generate the probability of an edge existing between node v and w as $p_e = \alpha p_c e^{-dis_{vw}/\Omega}$. If v and w have the same label, p_c will be set to 0.5 and if not, p_c will be set to 0.1. The component $e^{-dis_{vw}/\Omega}$ is used to control the influence of the locations of nodes. If the value of Ω is small, the value of $e^{-dis_{vw}/\Omega}$ will be greatly influenced by the distance between v and w. We make the number of average degree around 15 by adjust the value of α.

Table 1 shows the accuracy of different algorithms on different generated networks. We can see when the Ω is less than 10, which means the building of connections is greatly influenced by the location of nodes, our two methods can achieve a similar or higher accuracy than the Clauset's method. With the increasing of the value of Ω, the accuracy of Clauset's method performance better than our methods. So we recommend to evaluate the influence of geographic information first as described in section 3.1 before applying our methods. We also records the geographic span and average internal degree in different cases. In all levels of influence (Ω) that the geographic location on network structure, the connection locality method have the smallest geographic span. The geographic span of the node similarity method is smaller than the Clauset's method but larger than the connection locality method.

4.2 Twitter and Gowalla Network

In the real world, the factors which can influence the network structure can be very complex. We now test the algorithms on the networks generated by some real world applications, the Twitter and the Gowalla network. Since we do not have a community label for the real world dataset, we only apply the geographic span and the average internal degree of the communities to evaluate the detection results.



(a) Geographic span(T) (b) Ave internal degree(T)

(c) Geographic span(G) (d) Ave internal degree(G)

Figure 2: Analyzing the community detection results of different methods on the Twitter and Gowalla Network.

In Figure 2(a), we demonstrate the geographic span of different sizes (number of nodes in the community) of communities. From this figure we can see that under the random case, the geographic span is much larger and increases quickly to 800 kilometers. The communities detected by Clauset's method has a smaller geographic span. It begin with 280 kilometers when the community size is 2 but increases quickly when the community size become larger. Finally, the geographic span fluctuates between 500 to 600 kilometers. The two methods proposed in this paper have the best performance on controlling the geographic span on communities. Although the geographic span increases quickly when the community size becomes larger, these two method can keep the span much smaller than Clauset's method and the random case, especially for the method with the Equation 4 as the modularity. The geographic spans in different sizes of communities are only half of Clauset's method. Compared with the Twitter network, the geographic information in Gowalla has greater influence on the network structure. From Figure 2(c), we can see that our methods have a strong effect on limiting the geographic span of communities. Both the two methods can keep the span around or less than 200 kilometers. Especially for the connection locality method, even when the community size is very large, it can still keep the geographic span in a small range.

Another important observation is that in the highly geographically influenced networks, our method can also improve the network tightness in the communities. From Figure 2(b) and 2(d), we can see that the performances of these algorithm are similar to the case on the Twitter network. The different is that in the Twitter network, the connection locality method performs worse than the other two methods. But on the Gowalla network, it performs much better. This phenomenon illustrates that on the high geographically influenced networks, our method can improve the quality of the detection results on both geographic span and the tightness inside communities.

5. CONCLUSION

In this paper, we propose a new community detection method that keep the communities in small range of areas while maintaining the connection closeness of the nodes in the communities. We analyzed two real datasets and found that they have different level of locality. We performed extensive experiments on both synthetic and real world datasets. Results show that the proposed method find communities with nodes distributing in a smaller area compared with the traditional methods and having the similar or higher tightness on network connections. In our future work, we would like to explore low cost community detection algorithm utilizing the property of locality of nodes in communities.

6. REFERENCES

[1] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[2] Michele Coscia, Fosca Giannotti, and Dino Pedreschi. A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining*, 4(5):512–546, 2011.

[3] Jianbin Huang, Heli Sun, Yaguang Liu, Qinbao Song, and Tim Weninger. Towards online multiresolution community detection in large-scale networks. *PloS one*, 6(8):e23829, 2011.

[4] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[5] Jukka-Pekka Onnela, Samuel Arbesman, Marta C González, Albert-László Barabási, and Nicholas A Christakis. Geographic constraints on social network groups. *PLoS one*, 6(4):e16939, 2011.

[6] Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, and Vito Latora. Distance matters: geo-social metrics for online social networks. In *Proceedings of the 3rd conference on Online social networks*, pages 8–8, 2010.