# Detecting geographically dispersed overlay communities using community networks

Madhushi Bandara
LIRNEasia
12 Balcombe Place
Colombo 08, Sri Lanka
madhushi@lirneasia.net

Dharshana Kasthurirathna
Sri Lanka Institute of Information Technology
New Kandy Rd
Malabe 10115, Sri Lanka
dharshana.k@sliit.lk

Danaja Maldeniya
LIRNEasia
12 Balcombe Place
Colombo 08, Sri Lanka
danaja@lirneasia.net

Mahendra Piraveenan
Complex Systems Research Group
Faculty of Engineering & IT
University of Sydney
New South Wales 2006, Australia
mahendrarajah.piraveenan@sydney.edu.au

## ABSTRACT

Community detection is an extremely useful technique in understanding the structure and function of a social network. Louvain algorithm, which is based on Newman-Girman modularity optimization technique, is extensively used as a computationally efficient method extract the communities in social networks. It has been suggested that the nodes that are in close geographical proximity have a higher tendency of forming communities. Variants of the Newman-Girman modularity measure such as dist-modularity try to normalize the effect of geographical proximity to extract geographically dispersed communities, at the expense of losing the information about the geographically proximate communities. In this work, we propose a method to extract geographically dispersed communities while preserving the information about the geographically proximate communities. We do that by analyzing the 'community network' of the underlying network, where the centroids of communities would be considered as the nodes of a network. We argue that the inter-community link strengths, which are normalized over the community sizes, may be used to identify and extract the 'overlay communities'. The overlay communities would have relatively higher link strengths, despite being relatively apart in their spatial distribution. We apply this method to the Gowalla online social network, which contains the geographical signatures of its users, and identify the overlay communities within it.

## CCS Concepts

**•Human-centered computing → Social networks; Social network analysis;**

## Keywords

Social network analysis, Community detection, Geographically distributed communities

## 1. INTRODUCTION

Topological analysis of social networks have gained prominence in recent years. With the advent of network science as a separate field, modeling and characterizing self-organizing networks have been applied in a plethora of domains, ranging from biological networks, financial networks to social networks[4, 19, 13, 12]. One of the most vital pieces of information that is embedded in a social network is its community structure[9]. Communities can be regarded as the sets of nodes that may have homogeneous features among a diverse set of attributes. In particular, the communities that are extracted using topological features are extremely useful in social network analysis and mining, as it is the most fundamental and objective form of communities that can be extracted from a social network. Identifying and extracting communities from a social network may be vital in myriad applications which involve social network analysis, such as modeling social influence, information spread, epidemic modeling and defense related applications, among others. Moreover, extracting communities using a formal methodology may help to understand the structure and the operation of the social network in concern.

Different community detection algorithms have been proposed and have been applied in multitude of applications. Among them are the minimum-cut method, hierarchical clustering and modularity maximization[16, 19, 14, 20, 8, 22, 23]. One of the most widely accepted methods of community detection is modularity maximization, where the modular behavior of a network is utilized to identify and extract communities in a network.

One of the key limitations of the modularity maximization in community detection is that it doesn't take into account the contribution of geographical proximity that is vital in forming communities. That is, the nodes that are in close spatial proximity may tend to form communities in comparison to the nodes that are geographically apart. Thus, the interactions and links among the nodes that are geographically apart should carry more significance compared to the

nodes that are in close proximity in extracting communities. In order to address this limitation, the dist-modularity measure has been recently proposed[24]. This particular measure attempts to normalize the strength of links formed among nodes over their geographical distance. Thus, the dist-modularity measure may be used to identify the communities that are geographically distributed.

However, the dist-modularity measure has two key limitations. It requires relatively high computational time due to its computational complexity. Also, by normalizing the effect of geographical proximity of the constituent nodes in extracting communities, it actually disregards the communities consisting of geographically proximate nodes, which are equally as important as the geographically dispersed communities. Thus, it may not be used to capture the geographically proximate communities that are strongly connected with each other, while being geographically apart. Such communities can be observed in real-world networks such as migrant worker community networks and terrorist networks[11, 18], where the communities formed by geographically proximate nodes may have strong links with similar communities that may be geographically apart. Identifying and extracting such communities may provide vital information that may not be apparent in modularity based community detection algorithms.

In this work, we suggest that observing the interconnections of communities extracted through modularity maximization, in other words, analyzing the 'community networks', may pave way to identify and extract the geographically distributed communities. In order to do this, we suggest that the centroids of communities may be used to form a network of communities, where the links among such communities may be used to identify geographically dispersed communities or 'overlay communities'. Detecting such overlay communities may have interesting applications in areas such as indirect marketing, information propagation modeling and defense and counter terrorism domains. By influencing one geographically proximate community in a group of geographically dispersed communities, it may be possible to influence a geographically apart, yet closely connected, other communities. Identifying such overlay communities may be used as a computationally efficient method to extract geographically distributed communities.

The rest of this paper is organized as follows. The Background section provides an overview of the different community detection methods that are derived from modularity optimization, including the dist-modularity measure. The methodology section describes the proposed method of detecting overlay communities and how it is applied to a real-world social network dataset to extract geographically distributed overlay communities. Afterwards, the results obtained from the experimental analysis is presented. Finally, the concluding remarks are presented along with a brief description on potential future work.

## 2. BACKGROUND

Network science emerged as a prominent field of science with the proposition of the scale-free network model[4, 19]. The study of network science has facilitated observing networks in diverse domains such as social networks, biological networks and financial networks. In the recent years, much interest is given to extracting communities of social networks. The community information may be vital in understanding the information flows and the structure of an existing social network. While multitude of community extraction techniques such as the Hierarchical Clustering method and minimum cut method have been proposed to extract community information from social networks; modularity maximization remains the most widely used technique of extracting communities of social networks.

Louvain algorithm is a heuristic method developed by Blondel et al.[6], that partitions a social network into communities while optimizing Newman-Girvan (N-G) modularity of the partition. Louvain algorithm improves on the computational time of the modularity optimization technique, which is originally an NP-hard problem. Newman-Girvan modularity is used to measure how densely the detected communities of the partition are connected, relative to connections between these communities [6, 24]. In other words, the Newman-Girvan modularity measure is the fraction of edges within communities in the observed network minus the expected value of that fraction in a null model, which serves as a reference and should characterize some features of the observed network. Eq. 1 defines the Newman-Girwan modularity measure, which is used in the Louvain algorithm.

Consider a network that is modeled as an undirected graph G = (V;E) where V is a set of nodes and E is a set of relationships among nodes. The variables $n$ and $m$ represent the cardinalities of $V$ and $E$ respectively. Each edge $(v_i; v_j)$ is assumed to have an associated weight $w_{ij}$. For a given node $v_i \epsilon V$ , $\eta_i = \{v_j | (v_i, v_j) \epsilon E \vee (v_j, v_i) \epsilon E\}$ and $k_i = |\eta_i|$. Accordingly, the modularity $M(C)$ of a given partition $C$ is given as;

$$M(C) = \frac{1}{2m} \sum_{c \epsilon C} \sum_{i,j \epsilon c} w_{ij} - P_{ij} \tag{1}$$

$$P_{ij} = \frac{k_i . k_j}{2m} \tag{2}$$

Here, $P_{ik}$ refers to the null model that is used as a reference model, where the edges of the network are rewired randomly while preserving the degree distribution.

Multitude of social networks, including online social networks incorporate location information of the nodes in the network, in addition to the nodes and relationships among them. One important aspect in geographically distributed social networks is that the nodes in close proximity have an inherent nature of connecting with each other[24]. Thus, the community detection algorithms should ideally take into account this feature and normalize the effect of proximity to identify the actual communities in a social network.

As a result, a subsequent modularity measure called dist-modularity [24] has been proposed to normalize the effect of geographical proximity. This measure tries to identify the geographically distributed communities with a distance decaying function, under the assumption that the nodes that are in close geographical proximity have a higher tendency of forming community structures. This is an important assumption that we too employ indirectly, in formulating the idea of overlay communities that are geographically distributed.

The Eq. 3 gives the formal definition of the dist-modularity function.

$$M_{dist}(C) = \frac{1}{2m} \sum_{c \epsilon C} \sum_{i,j \epsilon c} w_{ij} - P_{ij} \tag{3}$$

$$P_{ij} = \frac{\widehat{P_{ij}} + \widehat{P_{ji}}}{2} \tag{4}$$

$$\widehat{P_{ij}} = \frac{k_i k_j f(d(v_i, v_j))}{\sum_{v_q \epsilon V} k_q f(d(v_q, v_i))}; f : R^+ \rightarrow (0, 1] \tag{5}$$

Here, $f$ is the distance-decay function. The basic assumption in dist-modularity optimization is that each node exerts a field on the surrounding nodes, which is inversely proportional to the distance from it. Thus, the null model used in the dist-modularity calculation assumes that nodes which are closer based on the distance function are more likely to be connected. This is the same assumption that we'd be utilizing to propose the idea of overlay communities where the Newman-Girvan modularity is considered to be likely to extract communities of members within the same geographical proximity.

The distance decaying function used in dist-modularity measure may further be extended to a gravity model where the inherent node properties are taken into account to capture the heterogeneity of the nodes[7].

$$P_{ij} = N_i N_j f(d_{ij} | d_{ij} = d) \tag{6}$$

where $N_i$ captures the importance of the node.

Thus, the distance-decaying function may be modified to capture the node heterogeneity as:

$$f(d) = \frac{\sum_{i,j|d_{i,j}=d} A_{ij}}{\sum_{i,j|d_{i,j}=d} N_i N_j} \tag{7}$$

which is the weighted average of the probability for a link to exist at distance $d$.

While the dist-modularity measure and its variants attempt to normalize of effect of geographical proximity in extracting communities, another branch of modularity maximization techniques attempt to harness the spatial information to extract the communities based on their geographical closeness. Spatially-near Modularity[10], which correlates with the spatial proximity of nodes, is an example of this particular approach of modularity maximization. Another interesting application of the modularity optimization is where it is been applied to extract multilevel communities based on a 'similarity attribute'. This particular application works under the assumption that nodes with similar features have a higher probability of being connected to each other. While this particular measure is useful in extracting communities that share the same geographical space, it is not much useful in extracting communities that are geographically distributed[17].

Based on the existing literature, two main approaches can be observed in extracting communities with geographical constraints. One is to extract geographically dispersed communities by normalizing the effect of geographical proximity. Dist-modularity and its variations are used for this purpose. The other approach is to harness or exploit the geographical proximity of communities and purposely consider the spatial nearness in extracting the communities. While these two approaches seem to contradict each other, the communities in social networks may encompass both geographically proximate as well as geographically dispersed communities. Thus, we attempt to propose a method to extract both the geographically proximate as well as geographically dispersed communities in a complimentary fashion. In other words, we propose a method to extract the geographically distributed communities, based on the interconnections of geographically proximate communities.

# 3. METHODOLOGY

In order to resolve this apparent dilemma where the geographically distributed communities have to be extracted without losing the information on geographically proximate communities, we propose the concept of 'overlay-communities', quite similar to the idea of 'overlay-networks' in peer-to-peer computing[3]. The idea is to extract the communities using the Louvain algorithm and then connect the extracted communities with inter-community links assuming that the nodes that are in close proximity have a higher probability of being in the same community[24].

When connecting the communities, we consider the centroids of each community as the 'node' of the community network, in order to assign a geographical location to each community. Afterwards, the connections among the members in communities are aggregated into 'links'. This way, we can easily quantify the geographical alignment of each community along with their inter-community link strengths. The link strength of each link are then divided by the multiplication of the sizes of the communities that it connects, in order to normalize the effect of the heterogeneity of community sizes. Normalizing over community sizes would help to identify the communities that are geographically distributed and yet strongly connected with each other, irrespective of the sizes of the underlying geographically proximate communities.

The communities that are strongly connected over the community network are termed as 'overlay communities', within the context of this work. Based on the assumption that the nodes that are in close geographical proximity tend to form communities, we may argue that the communities that are in close geographical proximity may tend to form strong connections with each other. Thus, the most interesting overlay communities would be those which are strongly connected yet whose centroids are further apart. Extracting such overlay communities may reveal information about the geographically distributed communities in social networks that are not apparent and that cannot be identified using the existing community detection algorithms. The Algorithm 1 explains the proposed technique in detail.

The proposed algorithm has a time complexity of $O(nlogn)$, which is the same as in the Louvain algorithm[6], in comparison to the exponential time complexity of the dis-modularity[24] measure. In order to test the effectiveness of the proposed algorithm, it was applied to a real-world online social network that encompasses geographical information of the users. We used a dataset from the Gowlla online social network[15] which has the geographical signatures of the users included in it for this purpose. By applying the above algorithm to the Gowlla network, we could extract the geographically dis-

**Algorithm 1:** Extracting overlay communities using community networks

```
1 Extract the community set C using the Louvain
  method of N-G modularity optimization;
2 for each community c in the set of communities C do
3     Identify the centroid of each community based on
      geographical location of each node in the
      community ;
4     Assign the centroid as the node representing that
      particular community in the community network ;
5 for each community pair p in the set of communities C
  do
6     Compute the strength of the link connecting the
      community pair p by aggregating the connections
      among the nodes in community pair p ;
7     Normalize the link strengths by the community sizes
      by dividing the link strengths by the multiplication
      of community sizes of the community pair p ;
8 Identify the communities that are relatively further
  apart geographically yet have relatively higher link
  strengths as the 'overlay communities' ;
```

tributed communities by identifying the communities that are strongly connected while being geographically apart.

The Gowalla network data set has 196,591 connected users as nodes. Check-in details of only 107,092 users were available for analysis, from which the home locations could be derived. There were 1,900,654 edges connecting the 196,591 users indicating the friendship between them. There were 6,442,892 total check-in records. It was observed that a relatively large portion of users did not have check-in records, resulting in unknown home locations for them. Hence, only the users with known location information were considered for the location analysis. We made the assumption that the other members were in similar vicinity. We ignored communities where locations of all members were unknown.

To derive approximate home location of each user, most frequent location for check-in was calculated. Based on that, all the check-in locations greater than 95% of the distance from the most frequent location were filtered out, assuming they represent anomaly trips of the user. Then, for each user, the weighted center of gravity of his check-in locations was calculated. That was considered the home location for them.

To detect the primary communities, we opted for the results of Louvain modularity optimization algorithm as it is computationally efficient and yields better results in comparison to many existing techniques. When the Gowalla social network was processed into communities by Louvain multilevel community detection algorithm, 5 non-overlapping community levels were detected. At level 1 network was broken into 19,396 communities, 2875 at level 2, 1025 at level 3, 839 and 820 at level 4 and 5 respectively. Level 5 produced the maximum modularity value for the network.

To study how the resulting communities are dispersed geographically, we used home locations of community members to calculate the standard distance deviation, which calculates the centroid of the community (with respect to the dispersion of community members) as well as the radius and the area of dispersion. The results indicated that when the modularity increases, the radius and area of each constituent community decreases. This further supports the assumption that the nodes that are in close geographical proximity have a higher tendency of forming communities.

The link strengths of the 820 communities extracted were measured. There were 4138 links connecting the 820 communities. The link density was observed to be relatively low. We then normalized the link strengths over the community sizes of the community pairs connected by each link in order to remove the effect of the heterogeneity of the community sizes, which could otherwise invariably affect the strengths of the inter-community links. The next section presents some of the results obtained using the analysis performed on the extracted community network.

# 4. RESULTS

The Fig. 1 depicts the distribution of the communities over the community size, in the community network formed using the above methodology. Based on the figure, it is evident that the community network contains relatively few communities with large number of members while relatively high number of communities with relatively smaller number of members. This is characteristic of the scale-free model.

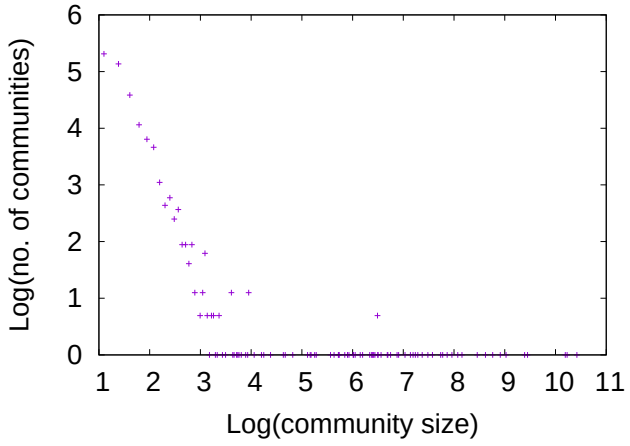



Figure 1: Distribution of communities over community size in logarithmic scale.

We further observe the degree distribution of communities in Fig. 2. According to the figure, the degree distribution fits well into a power-law degree distribution. The scale-free correlation and the scale-free exponent of the network were measured to be 0.74 and 0.67, respectively, further indicating that the community network fits into the scale-free model.

The Fig. 3 depicts a graphical representation of the community network obtained, where the link strengths were normalized over community sizes of the communities connected by each link. As the figure depicts, the community sizes and link strengths are heterogeneous and non-correlated in nature, suggesting that certain communities may be strongly connected, despite being geographically apart.

The Fig. 4[a] depicts that strength of each link against the Euclidean distance between the centroids of the communities that it connects. The link strengths are not nor-

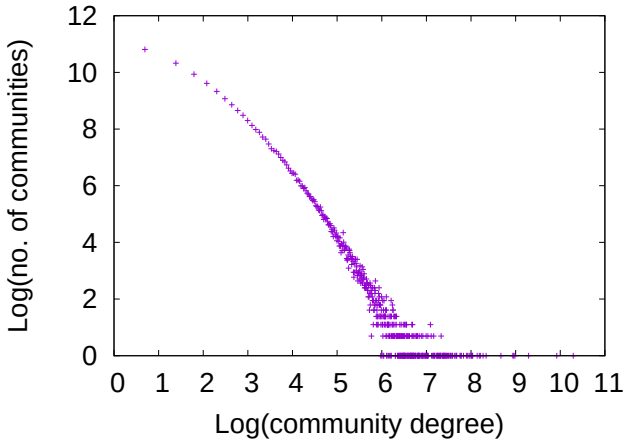



Figure 2: The degree distribution of the community network in logarithmic scale.

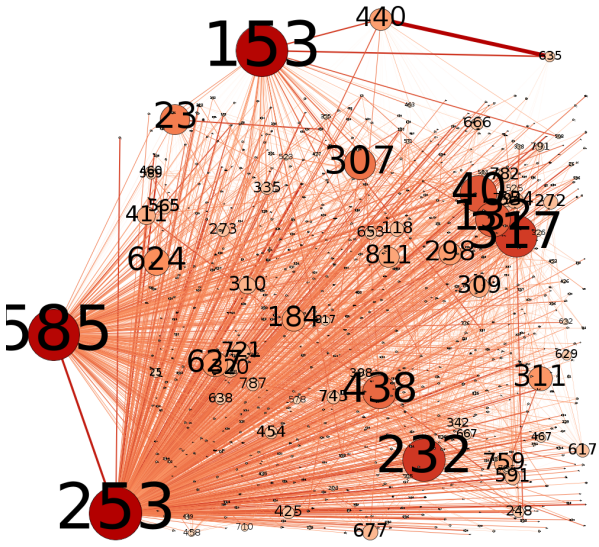



Figure 3: Community network with heterogeneous node sizes and normalized link strengths.

malized over community sizes. According to the figure, it is evident that there are certain links that show relatively higher strengths, while they connect communities that are geographically dispersed.

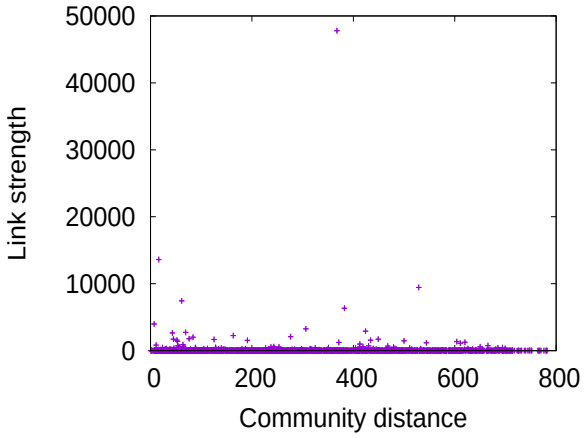
Fig. 4[b] shows the strength of each link, where the link strength of each link is normalized over the community sizes connecting them. Even after normalizing the link strengths, there are certain communities that are tightly connected despite being geographically apart. In the given figure, there is a link that particularly stands out in its link strength, although it connects two communities that are spatially apart. If we consider the link with the highest strength in the given distribution, it is the link connecting the community id 253 (population of size 26702) with the community with the id 585 (population of size 33560). The relative distance between these two communities is 368 units. When we normalize the tie strength by population, the link strength of these communities is higher than 82% of community pairs observed. Hence we cannot ignore this strong connection between the two communities or attribute it to being a result of the community size of one or both communities. It is important to note that these two communities are not overlapping and geographically apart significantly, when extracted using the Louvain algorithm, even though they have a strong connection between them. Thus, we could identify this particular community pair as a geographically dispersed single overlay community.

## 5. CONCLUSION AND FUTURE WORK

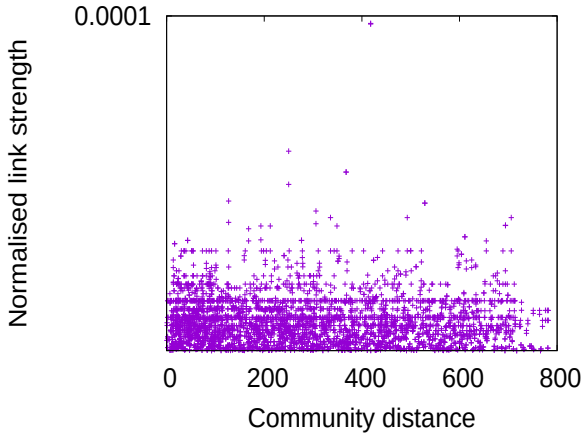
In this work, we propose a computationally efficient method to extract geographically dispersed communities, while preserving the information about the geographically proximate communities. We suggest that the Louvain algorithm may be used to identify the geographically proximate communities, which may then be connected by aggregating the connections among the nodes between each community. This is based on the assumption that the nodes that are in close geographical proximity have a higher tendency to form communities with each other. The centroids of each community would represent the nodes in the network of communities. These inter-community links could be used to form a 'community network'. The link strengths are then normalized over the sizes of the community pairs. In the resulting network, the community pairs that have relatively higher link strengths, while connecting relatively further communities are identified as geographically dispersed 'overlay communities'. These overlay communities may be used to identify geographically dispersed communities in applications such as migrant community detection and terrorist network detection, since the community networks in such scenarios have geographically proximate communities interacting with similar communities that are geographically apart.

Though we measure the Euclidean distance between two community centroids as the distance between communities, it may not be applicable for networks where communities are relatively large, making their centroids further apart, yet, overlapping at the periphery. Thus, non-overlapping and relatively smaller communities would be more appropriate to be considered for overlay community detection in our approach.

To our knowledge, there has not been an attempt to analyze a network of communities based on its topological properties. It may be possible to analyze the community




(a) Link strengths not normalized




(b) Link strengths normalized

Figure 4: The link strength among communities against distance between the community centroids. The link strength is determined based on the number of interconnections among communities. The distance between communities is measured in geographical coordinate based distance.

networks to extract more information about numerous network properties and behavior such as network resilience[5], assortativity[21], growth[4] and evolution.

Though we only consider community pairs in this work,the overlay communities could be in the form of sub-networks. Further, the overlay networks may be extracted at multiple levels of hierarchy. Thus, extracting these sub-networks and the hierarchical overlay networks could be the potential extensions of this work. While we consider the number of interactions within each community to denote a link and to measure link strengths, different network attributes may be used to form links and assign link strengths, resulting in networks of varying dimensions. Such community networks may be analyzed to extract information about the network that may not be revealed with existing network analysis techniques. Incorporating the extracted geographically distributed communities with Geographical Information Systems may reveal information about the patterns of interaction between communities that are geographically apart.

Though we consider only the networks that have a spatial bias in this work, such as friendship networks, there may be other networks, such as trust networks, where such a spatial bias may not be apparent. In such scenarios, the bias may occur due to some other dimension other than the spatial distance, such as the race or income level. In such scenarios, the communities that are further apart in that particular dimension but still are strongly connected could be identified using the proposed method, in order to reveal information about strong inter-community relationships that would otherwise may not be apparent.

## 6. ACKNOWLEDGMENT

The authors would like to thank Mr. Sriganesh Lokanathan, Prof. Rohan Samarajiva and Mr. Isuru Jayasooriya of *LIRNEasia*[2] for their support and contribution. The research was partly funded by the International Development Research Centre (IDRC)[1] of Canada.

## 7. REFERENCES

[1] International development research centre. https://www.idrc.ca.

[2] Lirneasia. http://www.lirneasia.net.

[3] K. Aberer, L. O. Alima, A. Ghodsi, S. Girdzijauskas, S. Haridi, and M. Hauswirth. The essence of p2p: a reference architecture for overlay networks. In *Peer-to-Peer Computing, 2005. P2P 2005. Fifth IEEE International Conference on*, pages 11–20. IEEE, 2005.

[4] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[5] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *nature*, 406(6794):378–382, 2000.

[6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[7] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*, 108(19):7663–7668, 2011.

[8] S. Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[9] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[10] J. Hannigan, G. Hernandez, R. M. Medina, P. Roos, and P. Shakarian. Mining for spatially-near communities in geo-located social networks. *arXiv preprint arXiv:1309.2900*, 2013.

[11] M. G. Herander and L. A. Saavedra. Exports and the structure of immigrant-based networks: the role of geographic proximity. *Review of Economics and Statistics*, 87(2):323–335, 2005.

[12] D. Kasthurirathna, A. Dong, M. Piraveenan, and I. Y. Tumer. The failure tolerance of mechatronic software systems to random and targeted attacks. In *ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pages V005T06A036–V005T06A036. American Society of Mechanical Engineers, 2013.

[13] D. Kasthurirathna and M. Piraveenan. Emergence of scale-free characteristics in socio-ecological systems with bounded rationality. *Scientific reports*, 5, 2015.

[14] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

[15] J. Leskovec and A. Krevl. {SNAP Datasets}:{Stanford} large network dataset collection. 2014.

[16] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th international conference on World Wide Web*, pages 695–704. ACM, 2008.

[17] X. Liu, T. Murata, and K. Wakita. Extracting the multilevel communities based on network structural and nonstructural information. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 191–192. International World Wide Web Conferences Steering Committee, 2013.

[18] R. Medina and G. Hepner. *Geospatial analysis of dynamic terrorist networks*. Springer, 2008.

[19] M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[20] M. E. Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[21] M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.

[22] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos. Community detection in social media. *Data Mining and Knowledge Discovery*, 24(3):515–554, 2012.

[23] M. Plantié and M. Crampes. Survey on social community detection. In *Social media retrieval*, pages 65–85. Springer, 2013.

[24] P. Shakarian, P. Roos, D. Callahan, and C. Kirk. Mining for geographically disperse communities in social networks by leveraging distance modularity. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1402–1409. ACM, 2013.