Extracting the Multilevel Communities Based on Network Structural and Nonstructural Information

Xin Liu [†‡§]
tsinllew@ai.cs.titech.ac.jp

Tsuyoshi Murata [†]
murata@cs.titech.ac.jp

Ken Wakita [†‡]
wakita@is.titech.ac.jp

[†]Tokyo Institute of Technology
2-12-1 Ookayama
Meguro, Tokyo
152-8552 Japan

[‡]CREST, JST
K's Gobancho, 7, Gobancho
Chiyoda, Tokyo
102-0076 Japan

[§]Wuhan University of Technology
122 Luoshi Road
Wuhan, Hubei
430070 China

ABSTRACT

Many real-world networks contain nonstructural information on nodes, such as the spatial coordinate of a location, profile of a person, or contents of a web page. In this paper, we propose Dist-Modularity, a unified modularity measure, which is useful in extracting the multilevel communities based on network structural and nonstructural information.

Categories and Subject Descriptors

G.2.2 [**Discrete Mathematics**]: Graph Theory—*graph algorithms, network problems*

Keywords

modularity; community structure; social network

1. INTRODUCTION

Modularity [2] is a measure for evaluating the "goodness" of a partition of a network into communities. The definition of modularity involves a comparison between the observed network and a null model, which serves as a reference. This null model should characterize some features of the observed network. However, the previously used null models are not good representations of real-world networks and thus result in less accurate modularity. A common feature of many real-world networks is "similarity attraction (SA)", i.e., nodes that are similar have a higher chance of getting connected. In this paper, we create a new null model that captures the SA feature. Based on this null model we propose Dist-Modularity. Compared with the famous NG-Modularity [2] proposed by Newman and Girvan, Dist-Modularity has the following advantages: 1) It applies to networks that contain nonstructural information. 2) It is useful in extracting the multilevel communities.

2. DIST-MODULARITY

For simplicity, we limit our vision to undirected networks. Suppose m and n are the numbers of edges and nodes, respectively. We use d_{ij} to denote the similarity distance between nodes v_i and v_j: the smaller of d_{ij}, the more similar of the two nodes. The estimation of d_{ij} is out of the focus

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

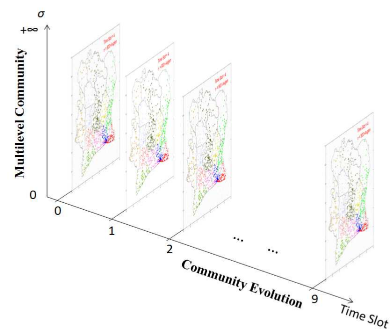


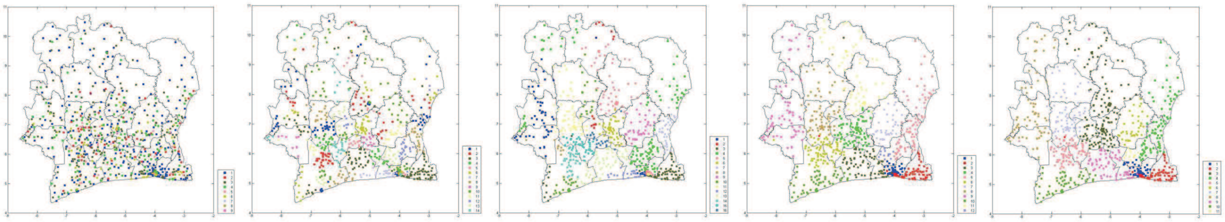
Figure 1: Extracting the multilevel communities and exploring the community evolution.

of this paper. In general, d_{ij} can be estimated by a distance function that takes the network structural or nonstructural information about v_i and v_j as input.

In the following, we first propose a new null model and then present Dist-Modularity. In our null model, the expected number of edges between v_i and v_j is defined as $P_{ij}^{\mathrm{Dist}} = \frac{\tilde{P}_{ij}+\tilde{P}_{ji}}{2}$, where $\tilde{P}_{ij} = \frac{N_i N_j f(d_{ij})}{\sum_{t=1}^{n} N_t f(d_{ti})}$. In this definition, we have a large freedom in specifying N_i and $f(d)$. N_i can be used for controlling the connectivity of v_i. To ensure that our null model preserve the number of edges of the observed network, N_i should satisfy the normalization condition $\sum_{i=1}^{n} N_i = 2m$. Beyond this condition, we can specify N_i freely. For example, N_i can be the degree k_i of v_i, or a representative attribute of v_i. $f(d)$ can be used to control the magnitude of the SA effect in our null model. For example, 1) if we specify $f(d)$ as a decreasing function, P_{ij}^{Dist} is negatively related to d_{ij}. Thus, nodes that are similar have a higher chance of getting connected — an evidence of the SA effect; 2) if we specify $f(d) = 1$, P_{ij}^{Dist} is not related to d_{ij}. Thus, the SA effect vanishes.

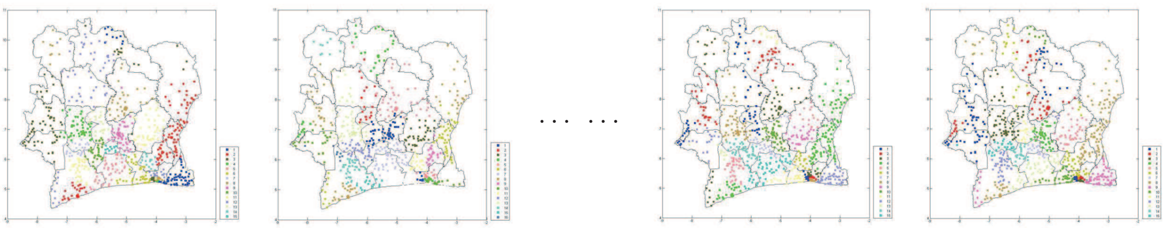
Based on the null model, we can define Dist-Modularity as $\mathrm{Q}^{\mathrm{Dist}} = \frac{1}{2m}\sum_{i,j=1}^{n}(A_{ij} - P_{ij}^{\mathrm{Dist}})\,\delta(l_i, l_j)$, where A_{ij} is the number of edges between v_i and v_j in the observed network, l_i is the community membership of v_i, and δ is the Kronecker's delta. Note that Dist-Modularity is a unified measure, since we can specify N_i and $f(d)$ freely and produce different $\mathrm{Q}^{\mathrm{Dist}}$. In particular, with $N_i = k_i$ and $f(d) = 1$, Dist-Modularity reduces to NG-Modularity. Besides, Dist-Modularity has the following advantages:

- It applies to networks that contain nonstructural information. Note that d_{ij} is at the heart of the definition



$\sigma = 0.1\,\bar{d}$ $\sigma = 0.5\,\bar{d}$ $\sigma = 1\,\bar{d}$ $\sigma = 5\,\bar{d}$ $\sigma = 10\,\bar{d}$

Figure 2: Extracting the multilevel communities along the σ axis.



Time slot = 0 Time slot = 1 . . . . . . Time slot = 8 Time slot = 9

Figure 3: Exploring the community evolution along the time slot.

of Dist-Modularity. In networks with nonstructural information, we can effectively use such information in the estimation of d_{ij}, and thus associate them with Dist-Modularity.

- It is useful in extracting the multilevel communities. We can generate Dist-Modularity by specifying $f(d) = \exp(-(d/\sigma)^2)$, where $\sigma \in (0, +\infty)$ is a parameter. By tuning σ we can adjust the decreasing rate of the function and thus the magnitude of the SA effect. Meanwhile, optimizing Dist-Modularity at different σ brings multilevel communities.

3. EXPERIMENT

To demonstrate the advantages of Dist-Modularity, we applied it to the antenna-to-antenna network of D4D dataset. This network is a spatial network where nodes and edges are embedded in space. It is based on records of mobile phone calls in Cote d'Ivoire. The nodes represent 1216 antennas which are associated with spatial coordinate information. The edges represent communications between antennas, with edge weight indicating the number of calls. Besides, this network is temporal: it has ten consecutive slices and each slice represents a two-week period record.

In spatial networks there is always "space effect", where long-range edges (i.e., the spatial distance between the two ends of the edge is long) are restricted due to cost. We are interested in the space-independent communities. That is, our goal is to take out the space effect and extract the hidden communities that are not due to the space factor [1]. Consequently, NG-Modularity fails to work, since it does not consider the spatial attribute of a node.

Note that the space effect is just our SA effect reflected in spatial networks: the two effects match when we estimate d_{ij} by the spatial distance between v_i and v_j. Thus we can simulate the space effect in the null model. Then, by comparing the observed network and the null model as the definition of Dist-Modularity, we are able to take out the space effect of the observed network and achieve our goal.

In specific, we specified $N_i = k_i$, $f(d) = \exp(-(d/\sigma)^2)$, estimated d_{ij} by the Euclidean distance of the coordinates of v_i and v_j, and employed Dist-Modularity optimization algorithm to this network. As shown in Fig. 1, we can extract the multilevel communities along the σ axis, and explore the community evolution along the time slot. Suppose $\bar{d} = \sum_{i,j=1}^{n} d_{ij}/n^2$ is the average distance of all node pairs. Fig. 2 shows the community structure in one of the network slice when σ equals to $0.1\bar{d}$, $0.5\bar{d}$, $1\bar{d}$, $5\bar{d}$, and $10\bar{d}$, respectively. Fig. 3 shows the community evolution at $\sigma = 1\bar{d}$. From Fig. 2 we can find that as σ increases, the community structure gradually correlates with the geography. In particular, the partition at $\sigma = 1\bar{d}$ matches the administrative subdivision of the country to a great extent. This example shows that Dist-Modularity successfully uses the network structural and nonstructural information for extracting the multilevel communities while NG-modularity fails.

4. CONCLUSION

We create a null model that captures the SA feature of real-world networks. Based on this null model we define Dist-Modularity, a unified modularity measure that incorporates NG-Modularity as a special case. Dist-Modularity is useful in extracting the multilevel communities based on network structural and nonstructural information.

5. ACKNOWLEDGMENTS

The authors are grateful to France Telecom-Orange for providing the D4D dataset. This work was partly funded by CREST JST and by NSFC under grant No. 61203154.

6. REFERENCES

[1] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte. Uncovering space-independent communities in spatial networks. *Proc. Natl. Acad. Sci. USA*, 108(19):7663–7668, 2011.

[2] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, 2004.