Map equation for link community

Youngdo Kim[1] and Hawoong Jeong[1, 2, *]

[1] *Department of Physics, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea.*
[2] *Institute for the BioCentury, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea.*

(Dated: August 15, 2011)

Community structure exists in many real-world networks and has been reported being related to several functional properties of the networks. The conventional approach was partitioning nodes into communities, while some recent studies start partitioning links instead of nodes to find overlapping communities of nodes efficiently. We extended the map equation method, which was originally developed for node communities, to find link communities in networks. This method is tested on various kinds of networks and compared with the metadata of the networks, and the results show that our method can identify the overlapping role of nodes effectively. The advantage of this method is that the node community scheme and link community scheme can be compared quantitatively by measuring the unknown information left in the networks besides the community structure. It can be used to decide quantitatively whether or not the link community scheme should be used instead of the node community scheme. Furthermore, this method can be easily extended to the directed and weighted networks since it is based on the random walk.

PACS numbers: 89.70.-a, 05.40.Fb, 02.10.Ox, 02.50.-r

I. INTRODUCTION

Complex networks have been widely used to represent the systems composed of connected objects, and many system-wide behaviors, which have emerged from the pattern of connections, have been successfully explained with the help of this simple model [1, 2]. The rising popularity of complex network through several disciplines—including statistical physics, computer science, computational biology, sociology, etc.—rests on many reasons; a major one is that many large scale networks have become available due to the advance in information technology. The advantage of the large-scale networks is that many meaningful statistical properties can be studied accurately, for example, degree distribution, clustering coefficient, assortativity, and motif profiles. However, the big size of the networks also brings disadvantages. When the network is small, it is very easy to visualize the network, and the organization structure of the network can be perceived intuitively. Instead, when the size becomes large, a comprehensive understanding of the structure could no longer be gained directly, and some quantitative analyses are required.

Community detection is one of the efforts devoted to the quantitative analysis of the organization structure. In many real-world networks, the nodes are connected neither regularly nor completely randomly. Instead, some nodes are densely inter-connected to form the communities, while these communities are loosely connected, relatively. This kind of network structure, which is usually referred as the community structure, is closely related to many dynamic processes on the network [3, 4]. Therefore, detecting the community structure has become one of the most important problems in the network research and many methods have been proposed to solve the problem efficiently [5]. The map equation method [6], also known as Infomap method, has been considered one of the best performing methods [5, 7]. This method is based on the Minimum Description Length (MDL) principle [8], according to which any regularity in the data can be used to compress the length of the data. Therefore, by considering the community structure as the regularity of the network and the path of the random walk on the network as the data to compress, the community structure can be detected during the compression of the path description. This is the main idea of the map equation method and it will be explained in detail in Sec. II

While most previous researches for community detection have focused on the community of nodes, some recent researches have started switching attention to community of links [7, 9] and even cliques [10]. From the theoretical point of view, the community of link could be more intuitive than the community of node in some real-world networks, because the link is more likely to have a unique identity while the node tends to have multiple identities. For example, most individuals in the society belong to multiple communities such as families, friends, and co-workers while the link between a pair of individuals usually exists for a dominant reason. From the practical point of view, overlapping communities of nodes, which is another attractive topic of community detection [11–14] could be detected as a byproduct because the links connected to a single node could belong to different link communities and consequently the node could be assigned to multiple communities of links. But exclusive partitioning of links is not always accurate and this problem is discussed in Sec. IV. The clique community is going further in this direction since a link is a clique of two nodes.

In this paper, we propose a modified version of the map

*Electronic address: hjeong@kaist.edu

equation method, which can be used to detect link communities under the MDL principle. In Sec. II, a brief review of original map equation is presented and the modified version of the map equation method is introduced in the following Sec. III. The best way to check the performance of a community detecting method is to compare the community result with the metadata available. We apply our method to several networks with rich metadata information, and the results are quantitatively compared with other methods for community detection in Sec. V. An important advantage of our method is that the results of link community and node community can be quantitatively compared. In Sec. VI, a model network is proposed to verify this property and the comparison is done in some real-world networks to show which partitioning scheme—link community or node community—can depict the organization structure of these networks more properly.

For the simplicity of derivation, only the binary and undirected network is considered in this paper. The extension to weighted and/or directed networks is briefly discussed at the end of Sec. III.

II. THE MAP EQUATION FOR NODE COMMUNITY

The most general definition of the community is that a community is a group of nodes that are densely interconnected. Meanwhile, from the viewpoint of information propagation, another definition can be proposed: A community is a group of nodes in which the information is more likely to be trapped rather than spread out. Considering that the random walk is the most fundamental model of information propagation, community structure can be detected by finding the local structure that traps the random walker. Some recent studies [15, 16] have showed that the modularity [17], which is a quality function used to find the communities as a group of densely connected nodes, can also be interpreted by the random walk and some disadvantages of the modularity can be easily resolved in this alternative approach.

The map equation method [6] detects communities by the information-propagation-based definition, under the philosophy of Minimum Description Length (MDL) principle [8]. The basic idea of the MDL principle is that any regularity in the data can be used to compress the length of the data. If we can find a way to encode the path of random walk on the network and consider the community structure as the regularity in the network, community structure can be detected by finding the partition that gives the minimum description length of the path. In the map equation method, the encoding rule for the path description can be described as follows.

To uniquely describe the path of a random walk on the network, the simplest way would be assigning a distinguishable code to each node in order to avoid the ambiguity, and the description length would become shorter when the more frequently visited nodes are given shorter code and less frequently nodes given a relatively longer code, which is the method known as the Huffman coding [18]. However, assigning a unique code to each node in the network could be very inefficient if the network size is large, and the movement of the random walker is frequently trapped in a small area—the community of nodes. A better strategy would be dividing the nodes into communities and using the codebook of two levels: The first level code describes the community that a node belongs to, and the second level code distinguishes a specific node from other nodes in the same community. In this strategy, a community (first level) code should be recorded in the path description when and only when the random walker enters the new community from other communities, and the random walks that is taking place within the community can be uniquely described by recording only the second level code. Additionally, an exit code should be assigned to each community, and it should be recorded when the random walker is exiting a community, so that the first level code and the second level codes can be distinguished. The costs of using the two-level codes would be fully compensated if the community structure is significant and it is well detected, because in this case the second level codes would become much shorter, and the first level codes of communities would not be frequently used, consequently reducing the total length of the path description. Therefore, the best partition of the network would be the partition that minimizes the average description length of the path of the random walk under the coding strategy described above.

Once the community partition $\boldsymbol{M}$ is decided, the probability of each code being used can be easily calculated and the map equation $L_{\text{nodecom}}(\boldsymbol{M})$, which is defined as the theoretical minimum of average description length, can be given by the Shannon's source coding theorem [19] as

$$L_{\text{nodecom}}(\boldsymbol{M}) = q_{\curvearrowright} H(Q) + \sum_{i=1}^{C} p^i_{\circlearrowright} H(P^i), \qquad (1)$$

where i is the index of community, α is the index of node, and C is the number of communities; $q_{\curvearrowright} \equiv \sum_{i=1}^{C} q^i_{\curvearrowright}$ is the total probability of using the first level codebook where $q^i_{\curvearrowright}$ is the probability of using the first level code for community i; $p^i_{\circlearrowright} \equiv q^i_{\curvearrowright} + \sum_{\alpha \in i} p_\alpha$ is the probability of using the second level codebook and the exit code for community i; and p_α is the probability of node α being visited, which is equal to the probability of using the second level code for node α. $H(Q)$ is the average description length contributed by the first level codebook:

$$H(Q) = -\sum_{i=1}^{C} \frac{q^i_{\curvearrowright}}{q_{\curvearrowright}} \log(\frac{q^i_{\curvearrowright}}{q_{\curvearrowright}}), \qquad (2)$$

while $H(P^i)$ is the description length contributed by the

second level codebook for community i:

$$H(P^i) = -\frac{q^i_{\curvearrowright}}{q^i_{\curvearrowright} + \sum_{\alpha\in i} p^i_\alpha} \log(\frac{q^i_{\curvearrowright}}{q^i_{\curvearrowright} + \sum_{\alpha\in i} p^i_\alpha}) - \sum_{\alpha\in i} \frac{p^i_\alpha}{q^i_{\curvearrowright} + \sum_{\beta\in i} p^i_\beta} \log(\frac{p^i_\alpha}{q^i_{\curvearrowright} + \sum_{\beta\in i} p^i_\beta}), \quad (3)$$

where p^i_α is equal to p_α when node α belongs to community i, otherwise zero. The probability $q^i_{\curvearrowright}$ is included in Eq. (3) to represent the contribution of exit codes for community i, and it can be computed from the following equation once the community structure $\boldsymbol{M}$ is given:

$$q^i_{\curvearrowright} = \sum_{\alpha\in i}\sum_{\beta\notin i} p_\alpha \frac{A_{\alpha\beta}}{k_\alpha}, \quad (4)$$

where $A_{\alpha\beta}$ is the element of the adjacency matrix, and it equals one if there is a link between node α and β, otherwise zero; $k_\alpha \equiv \sum_\beta A_{\alpha\beta}$ is the degree of node α. The description length is measured in bits if the logarithm is taken with base 2 in the equations above.

The community structure can be detected by finding the partition of nodes that minimizes the map equation $L_{\text{nodecom}}(\boldsymbol{M})$ in Eq. (1), just like other community detection methods based on maximization (or minimization) of the quality function. For example, many algorithms developed to maximize the modularity [17] can be directly used to minimize the map equation by replacing only the definition of quality function in the algorithms.

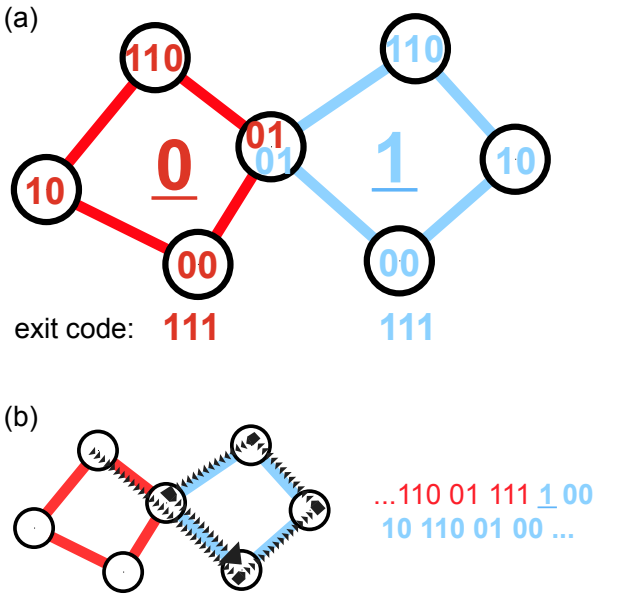

FIG. 1: (Color online) The encoding rule for the random walk path description in our method. (a) The links are divided into two communities: the red (dark gray) one and the blue (light gray) one, and the underlined first level codes are assigned to each link community. The second-level codes are assigned to the nodes, and the node in the center is given two second-level codes because it belongs to both communities. (b) An example of the random walk path is depicted on the left side, and the description of the path is given on the right side. The underlined first-level code is recorded only when the random walker is moving across the communities, and it is omitted when the random walker is moving within the community.

III. THE MAP EQUATION FOR LINK COMMUNITY

Although various kinds of methods [5] have been developed to find the communities, most of them are limited to the community of nodes. Some recent studies [7, 9], in which the link community is studied instead of the node community, showed that if the focus is alternated from the nodes to the links, a better description of the community structure could be found. In many real-world networks, a node could belong to several communities at the same time, and this fact makes the node community scheme fail to describe the organization structure of the system properly. For example, a person can belong to several social groups at the same time and an interdisciplinary research can belong to several scientific fields. Meanwhile, a link between a pair of nodes usually exists for a dominant reason, and the overlaps of communities over links would be less likely to happen compared to the overlaps of communities over nodes. The immediate advantage of the link community is that it can be used to detect overlapping communities of nodes, which is another active field of community identification [11, 12]. Although a link belongs only to a specific community when the links are partitioned into communities, a node could belong to multiple communities because the links connected to the node could belong to different communities (i.e., the link communities are overlapping over the nodes). A similar discussion can be applied to the cliques [9], which are the subnetworks of fully connected nodes, and the link community can be considered as a special case of clique community since a link is a clique composed of two nodes.

In this section, we propose a modified version of the map equation that can be used to find the communities of links. Since the original map equation can only be applied for node community, the encoding rule for the path of random walk needs to be modified. As illustrated in Figure 1, the first step of this modification is to let the partition $\boldsymbol{M}$ describe the link community instead of node community. The links are partitioned into communities, and the first level code is assigned to each link community. Meanwhile, the second level codes are still assigned to the nodes. The advantage of this encoding rule will be discussed later in Sec. VI. Since some nodes could belong to multiple communities in this case, each of these overlapping nodes would be given multiple second level codes, as many as the number of communities the node belongs to. Once the first- and second-level codes are assigned according to the community structure we assume, the path description is given as: (i) at each step, the ran-

dom walker is moving from the source node to the target node, which means the random walker is moving over a selected link that connects the source and target nodes; (ii) if the link bypassing at current step belongs to a different community compared to the community that the link of previous step belongs to, the first level code for community is recorded before recording the second level code for the target node; (iii) if the links of the current step and the previous step belong to the same community, the first level code would be omitted and only the second level code for the target node is recorded; (iv) additionally, an exit code should be inserted before each first level code in order to distinguish the first level codes from the second level codes.

The nodes that belong to multiple communities have multiple second level codes and this redundancy is likely to increase the length of the path description. However, if the link community is more significant than the node community (i.e., many nodes belong to multiple communities), the redundancy can be compensated by reducing the frequency of using first level codes especially when the random walker visits the overlapping nodes and move back to the previous community.

Once the encoding rule is given as above, we can get the map equation for link community if we know about the probability of using each code, and this computation of each probability can be easily done with the help of LinkRank [16]. LinkRank $r_{\alpha\beta}$, which is the probability of the link $\alpha \to \beta$ being visited by the random walker in the stationary state, is a constant value equal to $1/2M$ in the undirected binary networks, where M is the number of links in the network. We use $r^i_{\alpha\beta}$ to represent the community partition $\boldsymbol{M}$: $r^i_{\alpha\beta}$ is equal to $r_{\alpha\beta}$ if the link between nodes α and β belongs to community i, otherwise zero. Given the probability of visiting each link, the probability of using a second level code for node α in the community i is

$$p^i_\alpha = \sum_\beta r^i_{\beta\alpha}, \tag{5}$$

and the probability of using the first level code for community i is $q^i_\curvearrowright = \sum_\alpha q^i_{\alpha\curvearrowright}$, where

$$q^i_{\alpha\curvearrowright} = p^i_\alpha(1 - \frac{\sum_\beta r^i_{\alpha\beta}}{p_\alpha}), \tag{6}$$

is the probability that the first level code being used after visiting node α. Here p_α is the probability of visiting node α and it satisfies $p_\alpha = \sum_i p^i_\alpha = k_\alpha/2M$, where k_α is the degree of node α.

Finally, the map equation for the link community can be given as

$$L_{\text{linkcom}}(\boldsymbol{M}) = q_\curvearrowright H(Q) + \sum_{i=1}^{C} p^i_\circlearrowright H(P^i), \tag{7}$$

where $q_\curvearrowright = \sum_{\alpha,i} q^i_{\alpha\curvearrowright}$ is the total probability of using first level codes, and $p^i_\circlearrowright = q^i_\curvearrowright + \sum_\alpha p^i_\alpha$ is the total probability of using second level codes and the exit codes. $H(Q)$ is the contribution of first level codes to the average description length, and it can be computed by

$$H(Q) = -\sum_{i=1}^{C} \frac{q^i_\curvearrowright}{q_\curvearrowright} \log(\frac{q^i_\curvearrowright}{q_\curvearrowright}). \tag{8}$$

Similarly, $H(P^i)$ is the contribution of second level codes in community i to the average description length, and it can be computed from the following equation

$$H(P^i) = -\frac{q^i_\curvearrowright}{p^i_\circlearrowright} \log \frac{q^i_\curvearrowright}{p^i_\circlearrowright} - \sum_\alpha \frac{p^i_\alpha}{p^i_\circlearrowright} \log \frac{p^i_\alpha}{p^i_\circlearrowright}. \tag{9}$$

Now this map equation for link community can be used as the quality function to find link communities, just like other quality functions of community detection. Thus, most of the algorithms developed for other quality functions can also be modified to minimize $L_{\text{linkcom}}(\boldsymbol{M})$ in Eq. (7). In this paper, we used a modified version of the algorithm developed by Rosvall and Bergstrom [20], which is an extended version of the Louvain method [21]. The difference between our optimizing algorithm and the Louvain method is that the links, instead of the nodes, are locally grouped together to find the minimum efficiently.

This method can be easily generalized to weighted networks, in which weight is assigned to each link, and directed networks, in which direction is assigned to each link. In the weighted networks, the LinkRank $r_{\alpha\beta}$ is no longer a constant value, and it is proportional to the weight $w_{\alpha\beta}$ of each link. The remaining processes would just be the same. In the directed networks, the LinkRank $r_{\alpha\beta}$ is a quantity related to the global structure of the network, and it can be computed by following the processes described in Ref. [16]. If the directed network is composed of only one strongly connected component (SCC), in which a directed path always exists between any two nodes in the network, the equations in this section can still be directly used. It is important to notice that the sequences of α and β in Eqs. (5) and (6) are different. In the directed networks composed of more than one SCCs, the situation becomes complicated because there would be more than one stationary values for LinkRank. Therefore, the random hopping should be included in the random walk, which is the same as adding all-to-all links of small weight to the network, to ensure the existence of only one stationary value for the LinkRank. Thus, the original network becomes a all-to-all connected network and a link exists between any pair of nodes. This would make the minimization of the map equation computationally expensive, because the number of links to be partitioned would grow significantly. One possible solution is considering the random hopping links only when computing the LinkRank values and then normalize the LinkRank after removing the links generated by random hopping, as previously shown in Ref. [22].

During the submission of this paper, another extension of the map equation for overlapping community is

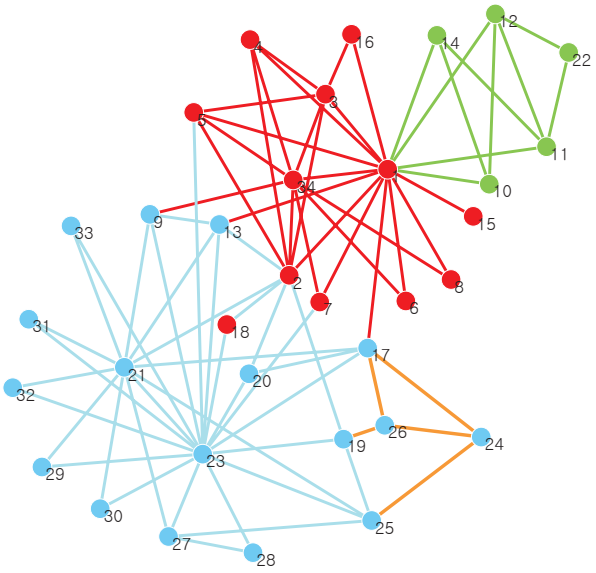

FIG. 2: (Color online) The communities detected in the karate club network. The color of the nodes indicate three node communities detected by the original map equation method, which is minimizing L_{nodecom}. The color of the links indicate four link communities detected by our method, which is minimizing L_{linkcom}. $L_{\text{linkcom}} = 4.28$ bits and $L_{\text{nodecom}} = 4.31$ bits, for the community results illustrated in the figure.

proposed by Esquivel and Rosvall [14]. It would be an interesting work to compare the performance of these two methods.

IV. A REAL-WORLD NETWORK ANALYSIS: THE KARATE CLUB NETWORK

We applied our method to the famous karate club network [23], a social network analyzed in most community detection researches, and the result is illustrated in Figure 2. The color of the links indicates the link communities detected by minimizing the map equation for link community, L_{linkcom}, while the color of the nodes indicates the node communities detected by minimizing the map equation for node community, L_{nodecom}. According to the result of node community, some nodes, especially nodes Nos.1 and 2, are categorized in the red (center) community, while a large portion of their neighbors belong to another community. In this karate club society, these nodes should be the members who connect different groups of people together, and their existence would be very important to integration of the whole society. However, the multiple social roles of these nodes are not captured in the node community scheme because the nodes are forced to belong to a single community. Meanwhile, the result of the link community, gives a much more intuitive interpretation of the organization structure. For example, the links connected to the No.2 node are divided into two communities, blue (lower left) community and red (center) community, and the red links are connecting other red nodes while most of the blue links are connected to the blue nodes. The links connected to node No.1 or to other nodes that are located at the boundary of communities, show similar behavior. The link community scheme properly describes the multiple roles of the overlapping nodes, and it gives a more intuitive organization structure than the node community scheme, at least in this example.

Meanwhile, it is important to notice that the link community approach is not the perfect solution to the detection of the overlapping communities. For example, nodes Nos.9 and13 should belong to both the red community and blue community at the same time according to the result of link communities, but the connection between those two nodes is categorized only to the blue community. This result may not represent the relation between those two members properly because the interaction between those two members very likely would be related to both the red and blue communities, not being limited to only one community as the link community result suggests. Thus, exclusive partitioning of the links may not represent the community structure of network well when communities of links highly overlap. However, the link community approach is a reasonable approximation that is quite effective in the practical applications. Firstly, its computational complexity is of the same level of the node community approach, while most other methods of detecting overlapping communities [11, 12, 14] require much more complex algorithms. Furthermore, the hard partitioning of links may not be an important issue if one is interested only in identifying the overlapping roles of the nodes because the degree of a node is usually larger than the number of the overlapping communities a node belongs to. For example, although the link between nodes Nos.9 and 13 is exclusively partitioned to the blue community, this result does not affect the detection of the overlapping roles of Nos.9 and 13.

V. COMMUNITY RESULTS COMPARED WITH METADATA

The qualitative explanation of the community detection results, although interesting, has its limits in verifying the validity of the methods. A more solid approach would be comparing the community results with the metadata contained in the system, like the analysis in Ref. [7]. We analyzed four networks with rich metadata, which are listed in Table I. The first is a sampled citation network of APS journal articles, which is constructed from the APS Data Sets for Research [24]. The sampled articles are the first- and second-level neighbors of a review paper [2] for complex networks. The metadata used to compare the results are the PACS (Physics and Astronomy Classification Scheme) numbers annotated to each article. Since the authors carefully choose the PACS

TABLE I: The real-world networks with metadata. N is the number of nodes, M is the number of links, C_{metadata} is the number of categories in the metadata, C_{linkcom} is the number of communities detected by our method, and C_{LC} is the number of communities detected by the link clustering method [7].

	N	M	C_{metadata}	C_{linkcom}	C_{LC}
APS sample	4755	29669	1076	339	14891
Metabolic [7]	1042	17512	169	156	2304
Philosopher [7]	1219	5972	5417	152	2777
Word Assoc. [7]	5018	55232	13141	765	36654

numbers to make their articles well publicized, it is reasonable to consider the PACS numbers as rich and trustful metadata. The other three networks were previously constructed and analyzed in Ref. [7]. The metabolic network is constructed from *E. coli* K-12 MG1655 strain, and the metadata used are the pathway annotations from the KEGG database [25]. The philosopher network is a network of Wikipedia pages for philosophers, with each link representing the hyperlinks in the articles, and the metadata are the categories that each page belongs to. The last network analyzed is the word association network, which is constructed from the datasets about free association of word pairs [26], and the metadata are the meanings or definitions assigned to each word in WordNet database [27].

In these networks, each node is annotated with single or multiple metadata, and the metadata can be considered as the overlapping communities because they are closely related to the grouping of nodes. Also, the result of our method, in which the communities of links are detected, could be considered as the overlapping communities of nodes. Thus, comparing the result of our method with the pre-assigned metadata can be considered as comparing two different results of overlapping communities. Although several criterions have been proposed for comparing overlapping communities, none of them is as conclusive as the variation of information (VI) [28], which is a well-defined and widely accepted criterion for comparing two *non-overlapping* community partitions. In order to overcome the disadvantage of individual criterion for overlapping community, we compare the metadata with the link community results by two fundamentally different criterions, the extended normalized mutual information (NMI) [29] and the extended Jaccard index [30], in order to observe the results from different aspects. Another extension of the mutual information for comparison of overlapping communities can be found in Ref. [14]. Although this method is a better approach compared to the extended NMI we used, it is not used in this work because in some of our examples one metadata may fully contain another metadata and the method cannot be used in this kind of cases.

The extended NMI is an information theory based

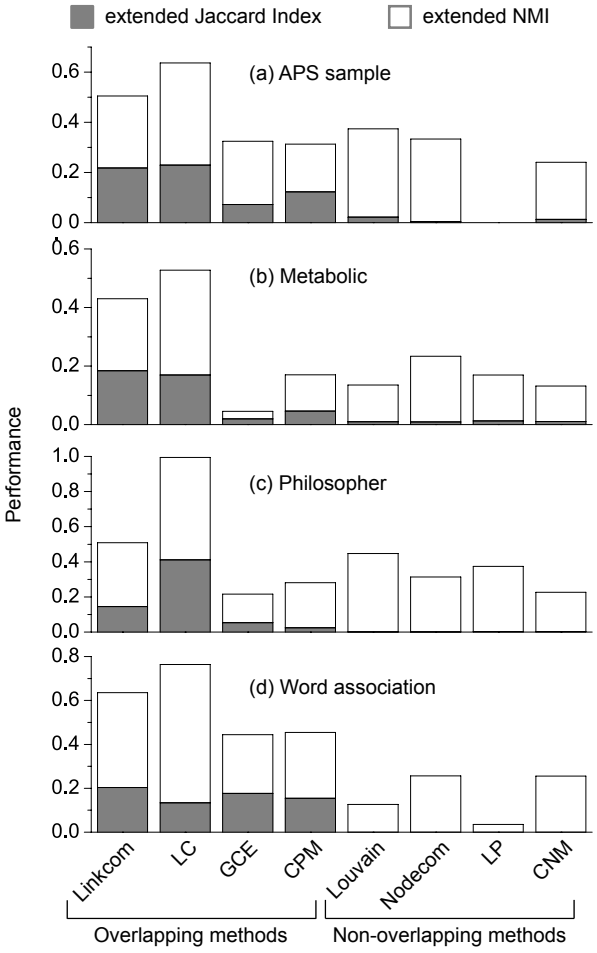

FIG. 3: The performance test of various community-detecting methods. The communities detected by each method are compared with the metadata, and the performance is measured by the extended Jaccard index and extended NMI. Linkcom represents the result of our method, LC represents the link clustering method [7], GCE represents the greedy clique expansion method [32], CPM represents the clique percolation method, Louvain represents the fast unfolding method in Ref. [21], Nodecom represents the original map equation method [6], LP represents the labal propagation method [33], and CNM represents the Clauset-Newman-Moore method [34]. The first four methods are able to detect overlapping communities, and the last four methods are not [35].

measurement and is defined as

$$N(\boldsymbol{X}|\boldsymbol{Y}) = 1 - \frac{1}{2}[H(\boldsymbol{X}|\boldsymbol{Y})_{\mathrm{norm}} + H(\boldsymbol{Y}|\boldsymbol{X})_{\mathrm{norm}}], \quad (10)$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are two different partitions of overlapping communities and $H(\boldsymbol{X}|\boldsymbol{Y})$ is the conditional entropy that measures the amount of information needed to infer $\boldsymbol{X}$ given the partition $\boldsymbol{Y}$. The extended NMI ranges from 0 to 1 and it equals to 1 only when two partitions $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identical. Meanwhile, the extended Jaccard coefficient falls into the category of external indexes that measure the similarity of two partitions statistically.

This index is defined as

$$\omega(\boldsymbol{X},\boldsymbol{Y}) = \frac{a_G}{a_G + d_G,} \tag{11}$$

where a_G and d_G measure the agreement and disagreement of partition $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively. The index satisfies $\omega(\boldsymbol{X},\boldsymbol{Y}) \in [0,1]$, reaching 1 only when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identical, and it reduces to the original Jaccard index in Ref. [31] when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are non-overlapping partitions.

We applied our methods to the four networks and the detected communities are compared with the metadata by the extended Jaccard index and the extended NMI. The result is presented in Figure 3, and the results of other community detection methods are also presented together to make a comparison. The first four methods, which are able to detect overlapping communities, show much better performance compared to the last four methods, which are able only to detect hard-partitioning communities. This result indicates the importance of detecting overlapping communities in recovering the properties of individual nodes. The first two methods, our method and link clustering method [7], which are detecting overlapping communities for nodes by detecting link communities, show significantly better performance—both the extended Jaccard index and the extended NMI showing meaningfully large value through the four networks analyzed—compared to other methods, indicating the overlapping communities for nodes can be efficiently detected by finding the link communities.

It is important to notice about our method and link clustering method, that both detect link communities but detect the communities at different hierarchical scales. As listed in Table I, the number of communities detected by the link clustering method is much larger than our method, indicating our method detects communities of relatively larger size and the link clustering method detects communities of relatively smaller size. It would be necessary to consider this scale factor when deciding which method to use in order to analyze the community structure of networks. It seems like this difference originates from the different optimization goal of two methods, but the fundamental cause of this difference is left unknown at this time.

VI. COMPARISON OF LINK COMMUNITY AND NODE COMMUNITY

It is interesting to notice that in the result of karate club network, which is illustrated in Figure 2, the map equation for link community, L_{linkcom}, is smaller than the map equation for node community, L_{nodecom}. Reminding that the map equation measures the amount of unknown information about the structure of the network assuming the community structure is already known, for each of L_{linkcom} and L_{nodecom} a smaller value of the map equation indicates that the community structure we assumed is a more proper description about the organization structure of the network. This reasoning can be extended to the comparison between L_{linkcom} and L_{nodecom}. In both of the methods, the rule for random walk is the same, the second level codes are all assigned to the nodes, and the description length is measured in the same unit. Therefore, the only possible cause for the difference between L_{linkcom} and L_{nodecom} is the different rules for the first level codes, and this difference can be used to test which encoding rule is better—the link community or the

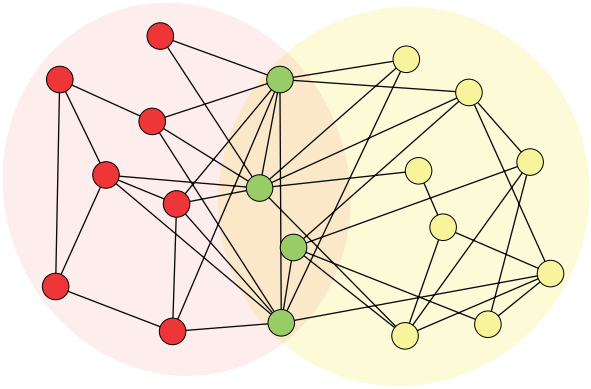

FIG. 4: (Color online) The model network that is proposed to verify the significance of overlap $\mathcal{O}$. This network is a variation of the Erdös-Rényi random network, and two communities, the red (left) and the blue (right), are embedded in the network. There are a total of $2N$ nodes in the network, and $2n$ of them (green or middle-gray nodes) are overlapping nodes while the other nodes are non-overlapping nodes, with $N - n$ nodes exclusively assigned to each community. The probability of connecting nodes in the same community is p_{in}, and it is much larger than p_{out}, which is the probability of connecting nodes from different communities.

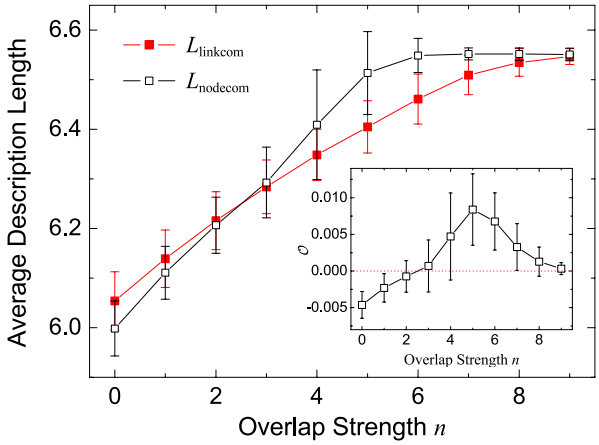

FIG. 5: (Color online) The minimum average description length detected by our method and the original map equation method in the model network. The red (filled) squares represent the minimum of L_{linkcom} given by our method, and the black (empty) squares represent the minimum of L_{nodecom} given by the original map equation method, for different values of overlap strength n. The inset shows the value of the significance of overlap $\mathcal{O}$.

TABLE II: The significance of overlap $\mathcal{O}$ measured in some real-world networks. The networks are listed by the descending order of $\mathcal{O}$.

Network	N	M	$< k >$	L_{linkcom}	L_{nodecom}	$\mathcal{O}$
High-energy theory collaborations [36]	8361	15751	3.8	5.89	6.56	0.0539
Network science collaborations [37]	1589	2742	3.5	3.48	3.77	0.0397
Power grid[38]	4941	6594	2.7	5.19	5.60	0.0380
Amazon.com co-purchase [7]	18142	46166	5.1	5.84	6.11	0.0224
Political blogs [39]	1490	19090	25.6	8.65	8.93	0.0163
Word association [7]	5018	55232	22.0	11.00	11.18	0.0080
APS journal citations (sampled) [24]	4755	29669	12.5	8.82	8.96	0.0076
Protein-protein interaction [7]	2729	12174	8.9	6.70	6.79	0.0068
Word adjacencies [37]	112	425	7.6	6.27	6.35	0.0068
Les Miserables [40]	77	254	6.6	4.64	4.68	0.0043
Political books [41]	105	441	8.4	5.44	5.48	0.0037
Zachary's karate club [23]	34	78	4.6	4.28	4.31	0.0035
Dolphin social network [42]	62	159	5.1	4.83	4.85	0.0024
Philosopher [7]	1219	5972	9.8	8.43	8.46	0.0018
Jazz musicians collaborations [43]	198	5484	55.4	6.91	6.91	0.0002
C. Elegans neural [38]	297	2359	15.9	7.52	7.46	-0.0041
E. coli metabolic [7]	1042	17512	33.6	8.33	8.25	-0.0053
American College football [44]	115	616	10.7	5.66	5.44	-0.0199

node community. For example, if the minimum value of L_{linkcom} is smaller than the minimum of L_{nodecom}, one can conclude that the link community scheme is better than the node community scheme in representing the organization structure of the network, because the link community scheme subtracted more information about the structure and left less unknown information in the path description. Instead, if the L_{nodecom} is smaller, this means that there is no much overlap of communities over the nodes and the non-overlapping methods are good enough to study the community structure of the network.

To quantitively analyze the difference between L_{nodecom} and L_{linkcom}, we propose a quantity called the *significance of overlap*:

$$\mathcal{O} = \frac{L_{\text{nodecom}} - L_{\text{linkcom}}}{L_{\text{nodecom}} + L_{\text{linkcom}}}. \tag{12}$$

This quantity measures how much better the link community scheme is compared to the node community scheme, and furthermore it can also be used to measure the overlapping strength of communities. The significance of overlap satisfies $\mathcal{O} \in (-1, 1)$, and it is positive when the link community scheme is better, being negative otherwise.

In order to check the validity of this quantity, we propose a model network (Figure 4) generated as follows. The model network is based on the Erdös-Rényi network [45], and two overlapping communities are embedded on the network. Among the $2N$ nodes of the network, $2n$ nodes are overlapping nodes, while $N - n$ nodes are exclusively assigned to each community. The probability of linking two nodes from the same community is p_{in}, and the probability of linking two nodes from different communities is p_{out}. The two communities overlaps more when n is larger and overlaps less when n is smaller. Therefore, n can be considered as the parameter that controls the overlap strength of the two communities. Figure 5 shows the results of L_{nodecom}, L_{linkcom} and the significance of overlap $\mathcal{O}$ for different values of n, while the set of parameters are fixed as $N = 50$, $< k >= 10$, $p_{out}/p_{in} = 15$. The error bar indicates the standard deviation over four hundred ensembles of the network realizations. When the overlap strength n is small, L_{nodecom} is much smaller than L_{linkcom}, indicating the node community scheme is better, and the significance of overlap $\mathcal{O}$ gets a negative value. As n grows, the significance of overlap $\mathcal{O}$ gets larger and it starts to get a positive value, which means the link community scheme is better in describing the organization structure. When n gets even larger, the overlap is too strong and the network is recognized as one community in both of the methods. Thus, the value of $\mathcal{O}$ falls to zero. This result matches our prediction well, therefore, the significance of overlap could be used as the quantitive measure of the strength of overlap.

We measured the significance of overlap $\mathcal{O}$ for some real-world networks and the results are listed in Table II. Although we do not fully understand how to interpret the exact value of $\mathcal{O}$ yet, some conclusions can still be made by comparing the values of $\mathcal{O}$ with the result of the karate club network (Figure 2), in which the overlap of communities is well observed. The significance of overlap in the karate club network is 0.0035, and many

networks show a larger value of $\mathcal{O}$ than the karate club network. Many social networks—such as the collaboration networks of scientists, the network of political blogs, the social network in *Les Miserables*, and the dolphin social network—show much stronger or similar degrees of community overlap compared to the karate club network, in accordance with the well-accepted knowledge that social communities tend to overlap with each other. The biological networks such as the *C. elegans* neural network and the metabolic network show negative values of $\mathcal{O}$, which means the communities in these networks do not overlap much, while the protein-protein interaction network shows a positive value of $\mathcal{O}$. This result might be related to the different biological functions of the communities in these biological networks, and further investigations would be necessary. The college football network, in which the teams are divided into regional leagues and most games happened within the leagues, shows a non-overlapping community structure and this result strengthens the validity of the significance of overlap. Finally, the fact that many networks have positive values of $\mathcal{O}$ indicates the overlapping community structure exists in many real-world networks, and it is important to study the organization structure of these networks by detecting the overlapping communities, instead of insisting on the non-overlapping communities.

VII. SUMMARY

We proposed a method to detect link communities in networks by modifying the map equation method, which detects communities by minimizing the average description length of the random walk. In our method, the communities are assigned to links instead of nodes, the encoding rule for the random walk is modified to represent this change in the community structure, and the corresponding map equation for the link community is proposed. The map equation for link community could be used to detect the link communities by finding the link partitioning that gives the minimum value of the map equation, just like other quality functions, and most of the algorithms that were developed to maximize (or minimize) other quality functions could be used after minor modifications.

One of the advantages of our methods is that the overlapping communities of nodes could be detected relatively easily, by defining the community of nodes by the communities of the links that are connected to the node. We tested our methods on some real-world networks by comparing the community results with the metadata of the nodes, and the result is compared with other community detection methods. The result shows that the communities detected by our method agree well with the metadata of the nodes, and the link community scheme is an efficient way to detect the overlapping communities of nodes.

Another important advantage of our methods is that the node community scheme and the link community scheme could be compared quantitatively. Since the difference between the map equation for the link community and the map equation for the node community comes only from the difference in community structure—the communities being assigned to the links or the nodes, the difference can be used to test which scheme, the link community or the node community, is better to represent the organization structure of the network. We used a quantity named as the significance of overlap to measure this difference in map equations, and the analysis of the significance of overlap in some real-world networks shows that many of the real-world networks can be better studied by the link community. Therefore, detecting the overlapping communities is necessary to understand the organization structure of the networks better, and finding link communities is an efficient way to detect the overlapping communities of nodes.

Acknowledgments

The authors thank Yong-Yeol Ahn and James P. Bagrow for providing us the network data along with the community results based on various community detecting methods. This work was supported by the Korean Systems Biology Research Project (20110002149) of the Ministry of Education, Science and Technology (MEST) through the National Research Foundation of Korea and by NAP of the Korean Research Council of Fundamental Science & Technology(KRCF).

[1] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[2] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[3] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani, Proc. Natl. Acad. Sci. U.S.A. **103**, 2015 (2006).
[4] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, Nat. Phys. **3**, 63 (2007).
[5] S. Fortunato, Phys. Rep. **486**, 75 (2010).
[6] M. Rosvall and C. T. Bergstrom, Proc. Natl. Acad. Sci. U.S.A. **105**, 1118 (2008).
[7] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, Nature **466**, 761 (2010).
[8] J. Rissanen, *An introduction to the MDL principle*, www.mdl-research.org (2004).
[9] T. S. Evans and R. Lambiotte, Phys. Rev. E **80**, 016105 (2009).
[10] T. S. Evans, J. Stat. Mech.-Theory E (2010), P12037 (2010).
[11] T. Nepusz, A. Petróczi, L. Négyessy, and F. Bazsó, Phys. Rev. E **77**, 016107 (2008).

[12] S. Gregory, J. Stat. Mech.-Theory E (2011), P02017 (2011).

[13] I. A. Kovacs, R. Palotai, M. S. Szalay, P. Csermely, PLoS ONE 5(9): e12528 (2010).

[14] A. V. Esquivel and M. Rosvall, e-print arXiv:1105.0812 (2011).

[15] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, Proc. Natl. Acad. Sci. U.S.A. **107**, 12755 (2010).

[16] Y. Kim, S.-W. Son, and H. Jeong, Phys. Rev. E **81**, 016103 (2010).

[17] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **103**, 8577 (2006).

[18] D. Huffman, P. IRE **40**, 1098 (1952).

[19] C. E. Shannon, ACM SIGMOBILE Mobile Computing and Communications Review **5**, 3 (2001).

[20] M. Rosvall and C. T. Bergstrom, *Fast stochastic and recursive search algorithm*, `http://www.tp.umu.se/~rosvall/algorithm.pdf` (2009).

[21] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech.-Theory E **2008**, P10008 (2008).

[22] M. Rosvall and C. T. Bergstrom, PLoS ONE 6(4): e18209 (2011).

[23] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1977).

[24] *APS Data Sets for Research*, `https://publish.aps.org/datasets`.

[25] M. Kanehisa, Nucleic Acids Res. **28**, 27 (2000).

[26] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber, Behav. Res. Meth. Ins. C. **36**, 402 (2004).

[27] C. Fellbaum, *WordNet: An Electronic Lexical Database* (The MIT Press, 1998).

[28] M. MEILA, J. Multivariate Anal. **98**, 873 (2007).

[29] A. Lancichinetti, S. Fortunato, and J. Kertész, New J. Phys. **11**, 033015 (2009).

[30] R. Campello, Pattern Recogn. Lett. **31**, 966 (2010).

[31] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data* (Prentice Hall College Division, England Cliffs, New Jersy, 1988).

[32] C. Lee, F. Reid, A. McDaid, and N. Hurley, e-print arXiv: 1002.1827 (2010).

[33] U. N. Raghavan, R. Albert, and S. Kumara, Phys. Rev. E **76**, 036106 (2007).

[34] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).

[35] The result of APS sample by LP method is missing from the figure, because the code for LP was not available.

[36] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).

[37] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).

[38] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).

[39] L. A. Adamic and N. Glance, *The political blogosphere and the 2004 US Election*, in Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem (2005)

[40] D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing* (Addison-Wesley, Reading, MA, 1993).

[41] `www.orgnet.com`.

[42] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, Behav. Ecol. Sociobiol. **54**, 396 (2003).

[43] P. M. Gleiser and L. Danon, Adv. Complex Syst. **6**, 565 (2003).

[44] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).

[45] P. Erdös and A. Rényi, Publ. Math. **6**, 290 (1959).