Understanding Topological Mesoscale Features in Community Mining

(Invited Paper)

Sue Moon† and Jinyoung You† and Haewoon Kwak† and Daniel Kim‡ and Hawoong Jeong‡
Department of Computer Science† Department of Physics‡
KAIST
335 Gwahangno, Yuseong-gu, Daejeon, Korea
sbmoon@kaist.edu, {jyyou, haewoon}@an.kaist.ac.kr, clearmind3927@kaist.ac.kr, hjeong@kaist.edu

***Abstract*—Community detection has been one of the major topics in complex network research. Recently, several greedy algorithms for networks of millions of nodes have been proposed, but one of their limitations is inconsistency of outcomes [1]. Kwak *et al.* propose an iterative reinforcing approach to eliminate inconsistency in detected communities.**

In this paper we delve into structural characteristics of communities identified by Kwak's method with 12 real networks. We find that about 40% of nodes are grouped into communities in an inconsistent way in Orkut and Cyworld. Interestingly, they are only two out of 12 networks whose community size distribution follow power-law. As a first step towards interpretation of communities, we use Guimera and Amaral's method [2] to classify nodes into seven classes based on the z-score and the participation coefficient. Using the z-P analysis, we identify the roles of nodes in Karate and Autonomous System (AS) networks and match them against known roles for evaluation. We apply topological mesoscale information to compare two AS produced by Oliveira *et al.* [3], and Dhamdhere and Dovrolis [4] We report that even though their AS graphs differ in size, their topological characteristics are very similar.

I. Introduction

Community detection has been one of the major topics in complex network research. Beyond the basic topological characteristics, the communities unveil the mesoscale structure of the network and provide insight into the topology. Quite a few community detection algorithms have been proposed in various areas from biology to computer science, but most of them have limited scalability due to their time complexity. Recently, several greedy algorithms have been proposed and have demonstrated to work for networks of millions of nodes [5], [6], [7]. Known limitations of modularity maximization approaches are resolution limit [8], its $\mathcal{NP}$-hardness [9] and inconsistency [1]. Although Brandes *et al.* state that "most suitable clustering algorithms probably identify one of [many intuitive clusterings structurally close]", inconsistent outcomes through multiple runs of one algorithm trouble researchers investigating individual communities.

Kwak *et al.* propose two quantitative metrics to assess the level of inconsistency among detected communities from independent multiple runs of modularity maximization methods [1]. They also propose an iterative reinforcing approach to eliminate inconsistency in detected communities. Their solutions successfully produce consistent communities in most networks, but show inconsistency in two huge networks with over 100 million links, namely, Orkut and Cyworld.

In this paper we delve into structural characteristics of communities identified by Kwak's method. Prior to interpretation of communities, we examine the percentage of nodes that undulate between communities and the community size distributions. We find that about 40% of nodes are grouped into communities in an inconsistent way in Orkut and Cyworld. Interestingly their community size distributions follow power-law, while others do not. It is yet to be shown whether the power-law distribution of community sizes is an artifact of modularity maximization algorithms or an inherent structure of a network.

As a first step towards interpretation of communities, we use Guimera and Amaral's method [2] to classify nodes into seven classes based on the z-score and the participation coefficient. The z-score and the participation coefficient represent how well a node connects other nodes in its own community and how all links of a node spread over neighbor communities, respectively. The seven classes broadly break nodes and hubs: ultra-peripheral, peripheral, non-hub connector, and non-hub kinless nodes; and provincial, connector, and kinless hubs. Using the z-P analysis, we identify the roles of nodes in Karate, C. Elegans, and Autonomous System (AS) networks and match them against known roles for evaluation. The AS graph represents the Internet connectivity at the service provider level and has been extensively studied [4], [3]. However, lack of complete knowledge has been a sour point in any measurement drive research of the AS Graph. In this work we use two AS graphs produced by Oliveira *et al.* [3] and Dhamdhere and Dovrolis [4] and compare their topological characteristics. We report that their AS graphs differ in size but their topological characteristics are very similar.

The remainder of this paper is organized as follows: In Section II, we review literature on community detection and then examine the size distribution of communities. In Section II, we discuss the identified communities as the topological mesoscale features. In Section III, we interpret the meaning of communities obtained from the AS graph. In Section IV, we conclude with discussion for future work.

978-1-4244-5489-1/10/ $26.00

II. COMMUNITIES: THEIR SIZES AND ROLES NODES PLAY

A. Review of community detection approaches

Community detection in complex networks has been an active topic of research in multiple disciplines. Communities represent a summary of structural features and functional grouping. Researchers have found community structures to be informative in human cellular signaling network [10], blogosphere [11], urban environment [12], and air transportation network [13].

Girvan and Newman have proposed a method to find community structure by betweenness centrality [14]. Their key idea is that intra-community edges have high edge betweenness. Their approach produce communities by removing edges of high edge betweenness. Radicchi *et al.* reduce the computing cost by considering only local quantities of edge-clustering coefficient rather than global quantities of edge betweenness [15]. The above algorithms are only two examples of many others from different disciplines. Clauset *et al.* put out a new metric *modulairty*, Q, defined as:

$$Q = \sum_i (e_{ii} - a_i^2) \tag{1}$$

where e_{ii} is the ratio of the number of intra-community links in the community i, and a_i is the ratio of links attached to nodes in community i [5]. It has been widely accepted and used in community detection. Greedy agglomerative algorithms are the fastest and most scalable solutions known so far [5], [6], [7]. However, they all produce inconsistent partitioning in multiple runs over the same network. In [1] Kwak *et al.* define two measures to assess consistency of partitioning over randomly ordered edge lists of a network: *pairwise membership probability* and *consistency*. Pairwise membership probability over N randomly ordered edge lists is defined like following:

$$p_{ij} = \frac{\sum_{n=1}^{N} \delta^n(c_i, c_j)}{N} \tag{2}$$

where

$$\delta^n(c_i, c_j) = \begin{cases} 1, & \text{if } c_i = c_j \text{ in the } n\text{th dataset} \\ 0, & \text{otherwise} \end{cases}$$

and i and j are adjacent nodes and c_i and c_j represent communities that i and j belong to, respectively. The p_{ij} is 0 if node i and j always fall in different communities over N runs, and the p_{ij} is 1 if node i and j always fall in the same community over N runs. The consistency, C, shows a network-wide level of consistency; it provides a summary of the pairwise membership probability of all edges. The consistency is defined as:

$$\mathcal{C} = \frac{\sum_{(v_i, v_j) \in E} (p_{ij} - 0.5)^2}{|E|} \times \frac{1}{(0.5)^2} \tag{3}$$

The consistency $\mathcal{C}$ weighs the pairwise membership probabilities away from 0.5. The second term in (3) normalizes $\mathcal{C}$ from 0 to 1. The consistency is 1 if the pairwise membership probability of every edge in a network is either 0 or 1, and the consistency is 0 if the pairwise membership probability of all edges is 0.

In [1], they report that the inconsistent partitioning can be produced even in a very small network, such as the Karate network, which consists of 34 nodes. They present an iterative solution to find a consistent (or robust as referred in [2]) communities in a network. They demonstrate that identified communities after 5 iterations (or cycles) from 10 real networks, are consistent and that their solutions produce very similar results among independent runs. However, for very large and dense networks with more than 100 million edges, namely, Orkut and CyWorld, their solution produce non-converging partitioning.

B. Analysis of community size distribution

In order to delve deeper into the converging communities of Kwak's method, we first look at what percentage of edges have pairwise membership probabilities that are not 0 or 1 after the 5th cycle. Here we use the 12 networks as in [1]. The 12 networks vary greatly in numbers of nodes and edges. In addition, they come from various fields: an off-line social networks (Karate), a biological network (C.Elegans), a protein interaction network (Protein Interaction), online bulletin board network (BBS), the Internet AS network (AS graph), online social networks (Facebook, Flickr, Orkut, YouTube, Cyworld), a sampled world-wide web network (WWW), and the Wikipedia link graph (Wikipedia). Table I summarizes the topological characteristics of the 12 networks.

In Figure 1 we plot the Cumulative Distribution Function (CDF) of pairwise membership probabilities after the 5th cycle, and observe that Figures 1(a) to (f) all have the pairwise membership probabilities of either 0 or 1, except for a very limited number of edges. That is, the edges either always belong to the same community or never after the 5th cycle. In Figures 1(g), (h) and (k) the number of edges with $0 < p_{ij} < 1$ is more than 100 but still they account for less than 1% of edges in the entire network. In the case of Flickr in Figure 1(i) $1,048,102$ or 4.611% of edges have $0 < p_{ij} < 1$, but the number is still not significant.

Community detection in Orkut and Cyworld by Kwak's method does not yield as high consistency as in other networks. Figures 1(j) and (l) show that in the two networks the percentage of edges that are still left undecided is large, both more than 40%; 41.23% for Orkut and 46.97% for Cyworld to be exact. First to look at is the number of runs per cycle. Both networks are too large and we cannot increase the number by an order of magnitude. Instead we add a few more cycles of 100 runs for these two networks and plot the CDF of the pairwise membership probabilities in Figure 2. The percentage of edges with $0 < p_{ij} < 1$ fluctuates from 40 even up to 60% in the case of Cyworld. The goal of this work is to identify the root of this divergent behavior.

The number of edges is known to have correlation to the level of consistency [1]. Then how about the characteristics of communities? How similar or different are the communities

Network	# of nodes	# of links	# of nodes in GCC		# of links in GCC		Avg. Degree	Link Density	Avg. C.C
Karate	34	78	34	(100%)	78	(100%)	4.6	0.14	0.57
C.Elegans	297	2,148	297	(100%)	2,148	(100%)	14.5	0.049	0.29
Protein	1,846	2,203	1,458	(78.9%)	1,948	(88.4%)	2.7	0.0018	0.071
BBS	7,410	103,462	7,339	(100%)	103,413	(100%)	28.2	0.0038	0.41
AS Graph	32,930	124,133	32,925	(100%)	124,131	(100%)	7.5	0.00023	0.38
Facebook	63,730	817,090	63,691	(99.5%)	816,886	(99.9%)	25.7	0.0004	0.22
WWW	325,729	1,090,108	325,729	(100%)	1,090,108	(100%)	6.7	0.000021	0.23
Wikipedia	1,870,709	36,532,531	1,870,521	(99.9%)	36,532,421	(99.9%)	39.1	0.000021	0.23
Flickr	2,302,924	22,838,276	2,173,369	(94.3%)	22,729,227	(99.5%)	20.9	0.00001	0.18
Orkut	3,072,440	117,185,083	3,072,440	(100%)	117,185,083	(100%)	76.3	0.000025	0.17
YouTube	3,223,588	9,376,594	3,216,082	(99.8%)	9,371,096	(99.9%)	5.8	0.000002	0.09
Cyworld	11,537,961	177,566,730	11,506,431	(99.7%)	177,548,838	(99.9%)	30.9	0.000003	0.16

TABLE I

BASIC STATISTICS OF 12 NETWORKS. GCC IS THE GIANT CONNECTED COMPONENT, AND AVERAGE C.C IS THE AVERAGE CLUSTERING COEFFICIENT [1].

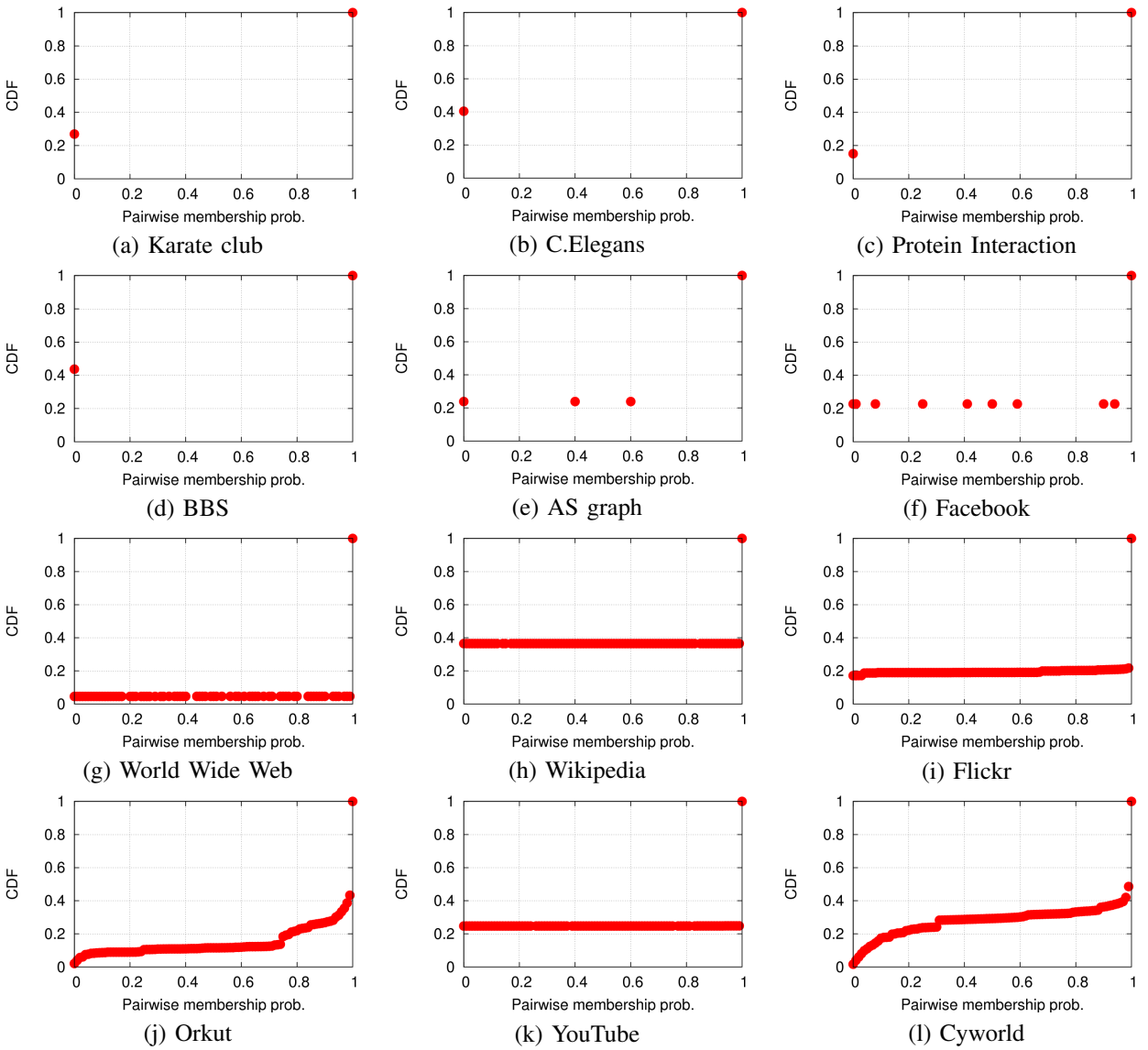
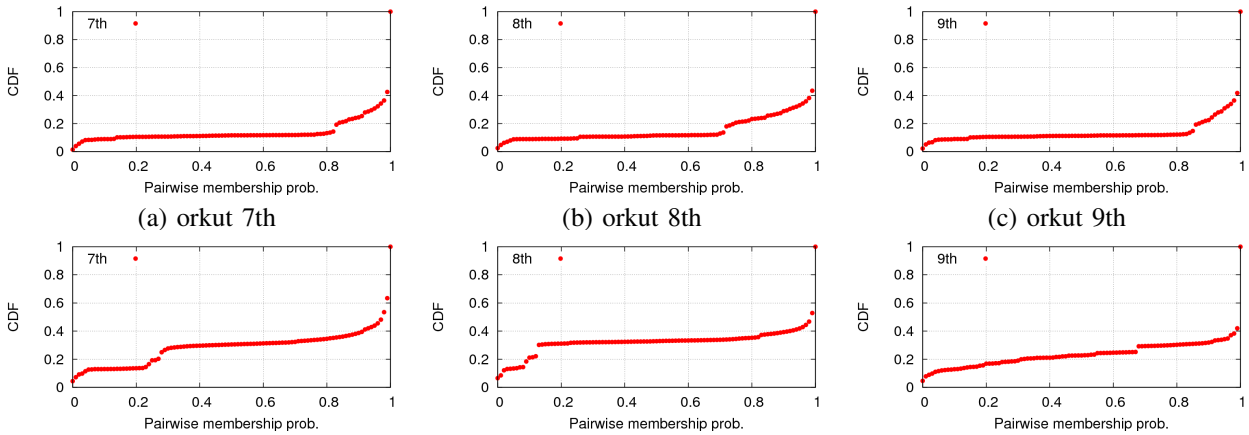


Fig. 1. Pairwise membership probability after the 5th cycle



(a) orkut 7th (b) orkut 8th (c) orkut 9th

(a) cyworld 7th (b) cyworld 8th (c) cyworld 9th

Fig. 2. CDF of pairwise membership probabilities of Orkut and Cyworld after 7th, 8th, and 9th cycles

from different runs? In order to answer these questions we examine the community size distribution of the 12 networks in Figure 3. For those networks with less than the convergence of 1 after the 5th cycle, we take the community partitioning solution with the highest modularity. We find a large number of small communities and a few large communities in Figure 3. For networks with more than 1 million nodes and 10 million edges, more than 70% of communities are smaller than 10 nodes. But very large communities that are only an order or two smaller in the number of nodes from the original network also emerge.

In order to investigate the tail behavior we plot the CCDF of community size distribution in Figure 4. The figures are labeld from (i) to (l) for ease of mapping to those in Figure 3. Only Orkut and Cyworld follow power-law in their community size distributions. Other studies have reported the power-law distribution of communities in various networks: the co-purchasing item network in Amazon [5], a Japanese online social network, mixi [16], and a scientific collaboration network [17]. Also other community identification methods lead to power-law distribution in community size [18], [19], [20].

We consider two possible explanations where the power-law in community size distribution comes from: the true nature of a network or the artifact of the algorithms. There have been much effort to analyze the characteristics of modularity [21], [22], and it is known that the size of communities grows with the size of the network in modularity maximization method [8]. Nevertheless, power-law distribution of community size has not been reported as an expected outcome of modularity maximization methods yet.

For all networks with more than $10,000$ nodes and $100,000$ edges we have investigated, top 10% communities hold more than 80% of nodes, except for BBS and Flickr. The BBS network is generated from user interactions in an online bulletin board system at a university with the population of about $7,000$ [23]. Each user has one's own bulletin board and write about personal things. Each bulletin board is considered a personal place. Anyone who can login to the bulletin board system can add comments to anyone else's board. Eom *et al.* construct the BBS network based on comments on individual boards [23] and the BBS network naturally reflect the offline intimacy between members. Therefore, an extremely large community seldom exists. The high average clustering coefficient of 0.48 confirms the prevalent modular structure in the BBS network [24].

C. Cartography

The pairwise membership probability and community size distributions have helped us understand the topological characteristics of complex networks. However, we have still not found out the roles that certain nodes play in communities. Guimera and Amaral define the role of a node in a network by two topological properties: *z-score* and *participation coefficient* [2]. The z-score z_i assesses how a node i connects to other nodes within the same community. A high intra-community degree leads to a high z-score. The z-score represents the authority of a node within the group. For example, two nodes of the highest z-score in Karate network are the two persons who became the boss in each group [25] after the breakup. The participation coefficient P_i measures how the links of a node i are well-distributed over neighbor communities. It is normalized between 0 and 14. If all links of node i are within the community, then $P_i = 0$. If the links of node i are well distributed over neighbor communities, P_i approaches 1.

Guimera and Amaral assume a normal distribution for the z-score and uses z_i of 2.5 as a cut-off point for hubs (top 1% nodes of high intra-community degree). They classify the other 99% nodes as non-hub. The hubs and non-hubs are further divided by the participation coefficient into three classes (R1, R2, and R3) and four classes (R4, R5, R6, and R7), respectively. Guimera and Amaral qualitatively label those seven classes: (R1) ultra-peripheral nodes; (R2) peripheral nodes; (R3) non-hub connector nodes; (R4) non-hub kinless nodes;

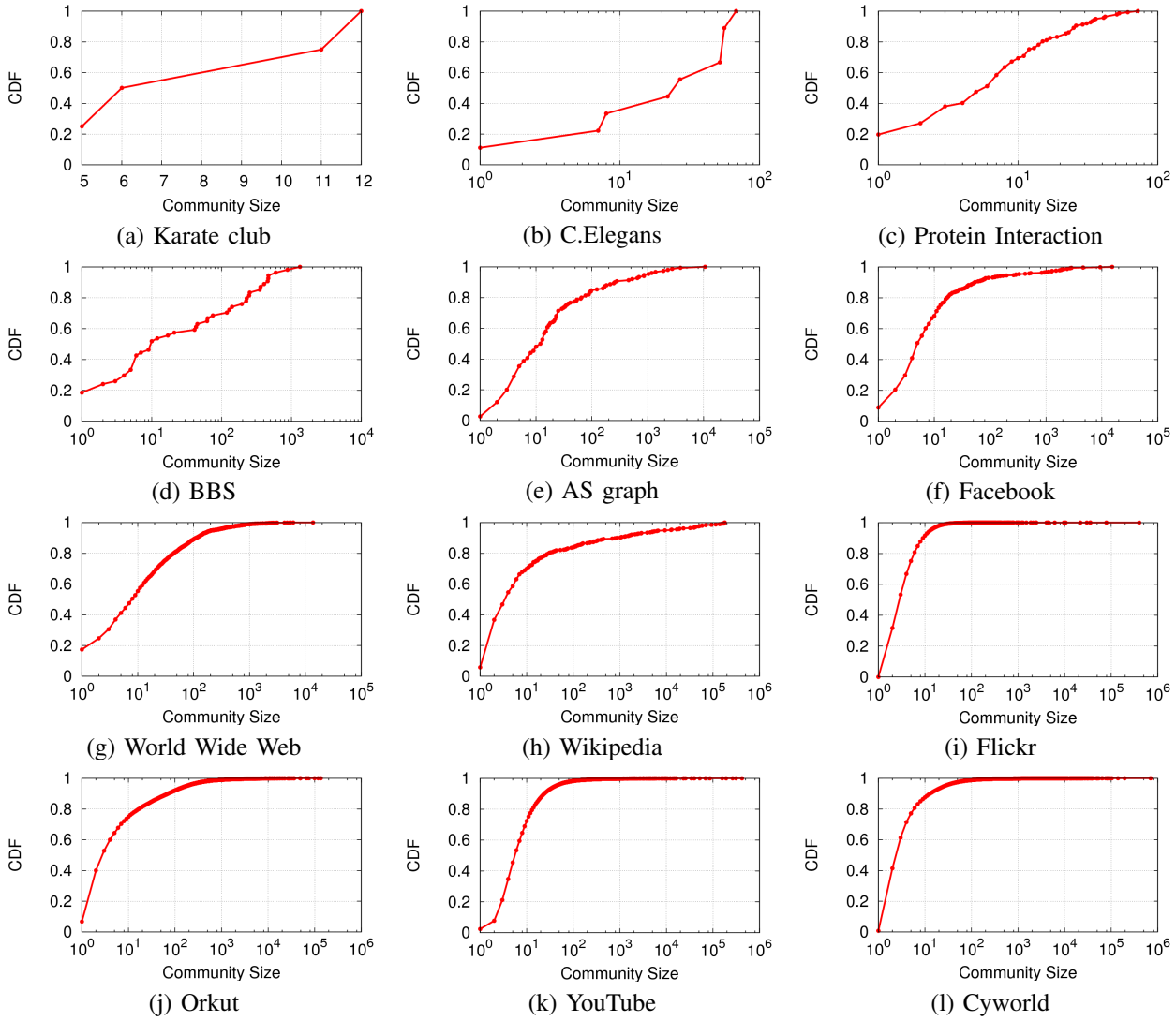


Fig. 3. Community size distribution

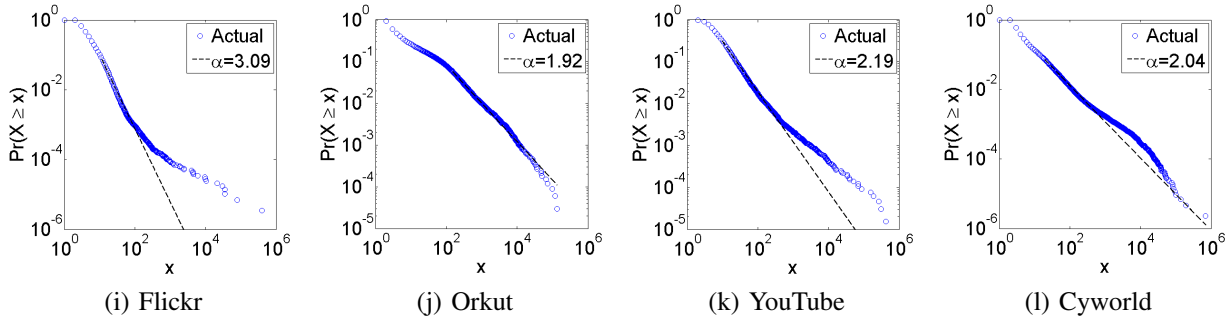
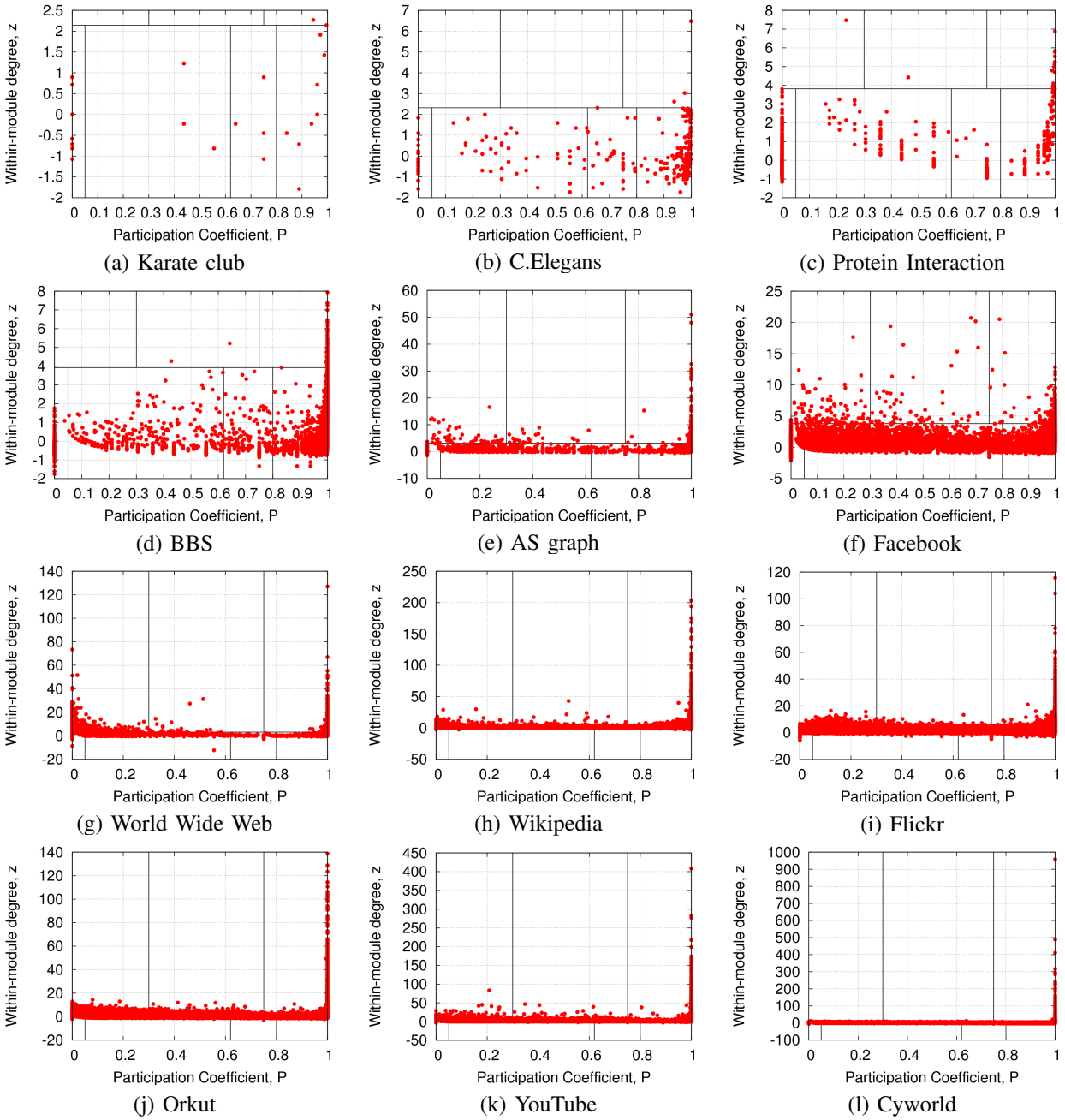


Fig. 4. Power law fitting

(R5) provincial hubs; (R6) connector hubs; and (R7) kinless hubs. We determine the pivot z-score between hubs and non-hubs as the 99% point from the empirical distribution of z-score, for no distribution of z-score obtained from 12 networks follows normal distribution.

In Figure 5 we illustrate the z-P space and divide nodes into seven classes by the z-score and participation coefficient. We find a pattern between the average degree of a network and the proportion of nodes falling in each class of the z-P plane. The lower the average degree of a network is, the more



(a) Karate club (b) C.Elegans (c) Protein Interaction

(d) BBS (e) AS graph (f) Facebook

(g) World Wide Web (h) Wikipedia (i) Flickr

(j) Orkut (k) YouTube (l) Cyworld

Fig. 5. z-P plane

nodes fall in R1. In Protein, WWW, and YouTube networks, we find 77.57%, 91.89%, and 71.98% of nodes fall in R1, respectively. The nodes of R1 are low in both the z-score and the participation coefficient; a node mostly connects to other nodes within its own community in a sparse network. In contrast, if an average degree of a network is higher, more nodes fall in R4. The nodes in R4 have low z-score and high participation coefficient. That is, links of a node is likely to spread evenly over neighbor communities in a dense network. These nodes that connect to many communities might lead inconsistency in community identification, because they have many choices in choosing their community membership. For future work we will investigate possible correlation between the unsettling edges in Figure 1 and nodes in R4 in Figure 5.

III. Validating Identified Communities

The basic assumption we make when using community detection algorithms is that the topological structure of a network presents the roles that nodes plays. In a typical human organization there is a hierarchy such that, the top manger talks to secondary managers, but unlikely to floor workers. Similar interpretations are possible in other types of networks.

Graph	z_{99}	R1	R2	R3	R4	R5	R6	R7	0/0
Karate club	2.14377	15	5	4	9	0	0	1	0
		44.1%	14.7%	11.8%	26.5%	0.0%	0.0%	2.9%	0.0%
C.Elegans	2.32009	39	52	31	172	0	0	3	0
		13.1%	17.5%	10.4%	57.9%	0.0%	0.0%	1.0%	0.0%
Protein Interaction	3.81447	1,131	92	78	142	1	1	13	0
		77.6%	6.3%	5.3%	9.7%	0.1%	0.1%	0.9%	0.0%
BBS	3.92132	2,320	602	450	3,893	0	2	72	0
		31.6%	8.2%	6.1%	53.0%	0.0%	0.0%	1.0%	0.0%
AS graph	3.12862	22,525	2,545	3,199	4,321	40	27	263	6
		68.4%	7.7%	9.7%	13.1%	0.1%	0.1%	0.8%	0.0%
Facebook	3.78518	19,915	11,328	4,266	27,231	113	60	461	18
		31.4%	17.9%	6.7%	43.0%	0.2%	0.1%	0.7%	0.0%
World Wide Web	3.07442	296,728	8,334	2,389	12,248	2,157	29	1,044	2,800
		91.1%	2.6%	0.7%	3.8%	0.7%	0.0%	0.3%	0.9%
Wikipedia	1.86745	365,518	266,435	102,693	1,117,164	815	341	17,549	6
		19.5%	14.2%	5.5%	59.7%	0.0%	0.0%	0.9%	0.0%
Flickr	3.60555	1,202,898	101,795	178,702	474,620	2,028	1,089	15,325	196,913
		55.3%	4.7%	8.2%	21.8%	0.1%	0.1%	0.7%	9.1%
Orkut	3.38152	900,393	323,524	62,296	1,755,560	2,337	306	28,003	22
		29.3%	10.5%	2.0%	57.1%	0.1%	0.0%	0.9%	0.0%
YouTube	2.10732	2,313,546	202,319	214,889	451,374	9,642	2,732	19,760	1,821
		71.9%	6.3%	6.7%	14.0%	0.3%	0.1%	0.6%	0.1%
Cyworld	2.93061	2,225,262	1,240,314	495,064	5,682,655	9,589	2,351	84,943	180,858
		22.4%	12.5%	5.0%	57.3%	0.1%	0.0%	0.9%	1.8%

TABLE II
NUMBER PERCENTAGE OF NODES IN EACH CLASS OF z-P PLANE

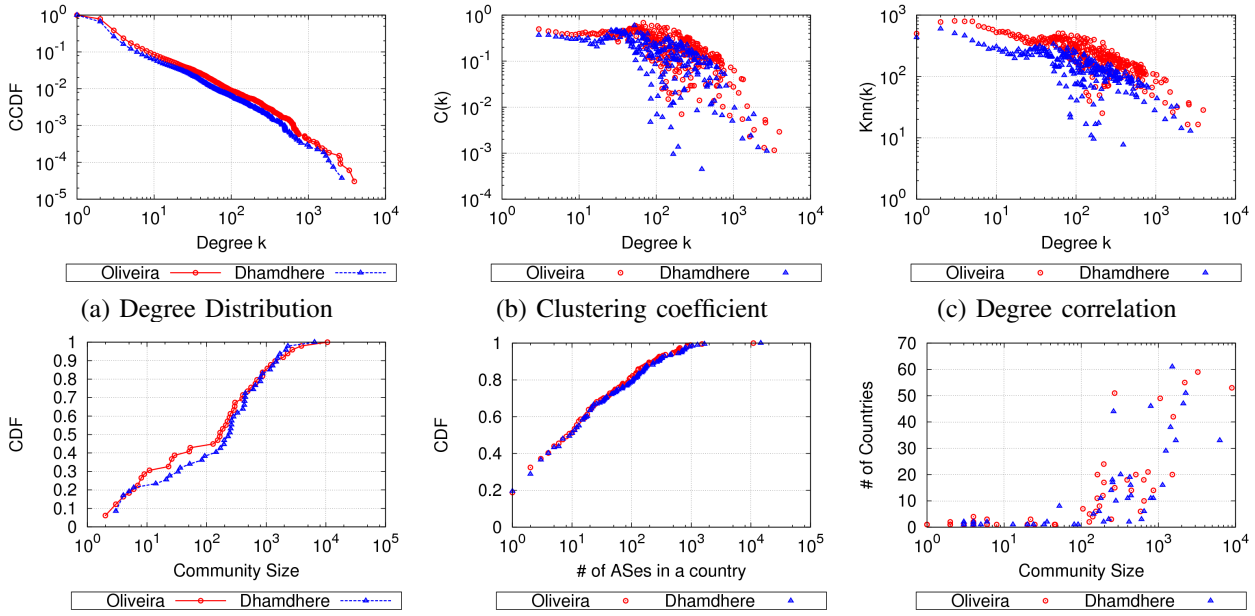
Meunier *et al.* have acquired functional Magnetic Resonance Imaging (fMRI) data of 18 volunteers and applied the Louvain method [7] in order to discover the hierarchical modular structure of human brain networks [26]. Their findings of nodal structures turn out to match well-defined neuroanatomical systems, but the results are yet empirical and require further validation. Sociologists and psychologists are one of the first to have built the human social network. Stanley Milgram's famous six-degrees-of-separation experiment has brought us the surprising insight that we live in a small world [27]. In his work on the strength of weak ties Granovetter defines a link based on committee involvement and discovers the power of inter-cluster links [28]. Sociologists construct networks from corporate financial data, membership to boards of directors, all sorts and analyze the structure of the network. However, sociologists' field work involve data collection via survey and their networks have been traditionally limited to the order of thousands. The collaboration network generated by published papers reflects the relationships among authors. Rodriguez and Pepe detect communities in a coauthorship network created from 560 manuscripts of sensor networks and wireless communication [29]. They find that identified communities reflect academic department and affiliation of individuals, but not the country of origin and academic positions of individuals. For the case of biological networks, Ravasz *et al.* succeed to identify functional modules hidden in the metabolic network of 43 different organisms [30]. They find that those modules are connected in a hierarchical manner, such that many small, highly connected modules are organized into larger, less cohesive units. In [31] protein-protein interaction network has also been analyzed and shown that identified modules, densely connected within themselves but sparsely connected to the rest of the network, are actual protein complexes and dynamic functional units.

Kwak *et al.* demonstrate in their preliminary work the impact of geographic relevance in detected communities in the AS graph [1]. They report on three communities: the largest community, a single-country-AS dominated community, and a star-shaped community. In this section we conduct a more comprehensive study to validate the meaning of communities in the AS graph.

A. AS graph

In the explosive growth path of the Internet, the Internet AS topology has evolved in a distributed manner. There is neither a central policy body to determine the direction of evolution nor an administrator to monitor and provision for better performance of the Internet and to plan the development strategy for the entire Internet. The distributed nature of growth gives the Internet not only the scalability but also the uncertainty; it is a challenge to obtain the complete topology of the Internet AS graph. Worse yet, the *observed* topology in the AS graph reflects only the existence of connectivity between two ASes; not the actual number of links, the capacity of links, or traffic volume. Moreover, the topology of the collected AS graph largely depends on the observation method and the resulting topological characteristics may differ.

Faloutsos *et al.* report simple power-laws hold in the Internet AS topology [32]. Later Chen *et al.* show that the degree distribution of AS topology does not follow strict power-law but heavy-tailed [33]. Carmi *et al.* investigate the structure of Internet AS graph by the k-shell decomposition method [34]. They divide the Internet into three components:



(a) Degree Distribution (b) Clustering coefficient (c) Degree correlation

(d) Community Size (e) No. of ASes (f) No. of Countries

Fig. 6. AS graphs by Dhamdhere and Oliveira

a nucleus, a fractal component, and dendrite-like structures. Their findings are similar with the Internet Jellyfish model proposed by Siganos *et al.* [35]. Mahadevan *et al.* analyze the properties of Internet topology by degree correlation in d-sized subgraphs [36]. They show that the Internet AS topology can be reconstructed with the degree correlation when $d = 3$.

Oliveira *et al.* proposes a model that explains the topology dynamics as a consistent-rate birth and death process [3]. Its model identifies topology changes in observed data, and its inference is statistically significant. Dhamdhere and Dovrolis observe linear growth trend in terms of ASes over ten years [4], while Magoni and Pansiot examine the exponential growth during 1997-2000 [37].

We analyze the community structure of two AS graphs and compare them. One is produced by Dhamdhere and Dovrolis (which we refer simply as (D)) [4] and the other is by Oliveira *et al.* [3] (later referred to as (O)). Dhamdhere's AS graph contains $26,831$ ASes in 191 countries and $72,985$ links between the ASes. Oliveira's AS graph is larger than Dhamdhere's; it contains $32,930$ ASes in 191 countries and $124,133$ links between ASes.

First, we compare the basic topological characteristics of the two AS graphs. Figure 6(a) shows the degree distribution of two AS graphs. We present that their degree distributions follow power-law; the exponents are 2.15 (D) and 1.93 (O). We obtain similarly close values of clustering coefficients and assortativity (Figures 6(b) and (c)). The analysis of these three topological characteristics support that two AS graphs are similar structurally.

Using Kwak's consistent community identification method [1], we identify 47 communities in (D) and 49 in (O). We plot the size distribution of communities in Figure 6(d). Note that there is a difference of about $5,500$ in the number of nodes between two graphs and thus their communication size distributions cannot be perfectly aligned.

The first question we raise about the communities in the AS graph is whether all the nodes in a community are from a single country or not. Are the communities geographically bounded? Dhamdhere and Dovrolis publish the location of the company managing each AS by using WHOIS queries [4]. We use the country code from their dataset and map ASes to countries.

In Figure 6(c), we plot the number of ASes in a country. The 191 countries have at least one AS. We find that the geographical location of ASes is highly skewed. By the number of ASes, the top ranked country is the United States. It contains $11,084$ ASes in (D) and $11,019$ ASes in (O). Considering that the 2nd ranked country of Russia has only $1,528$ ASes in (D) and $1,512$ in ASes (O), the concentration of ASes ni United States is significant. Top 10% of countries contain 88% of ASes, and top 20% of countries contain 95% of ASes. On the other hand, the proportion of countries containing only a single AS is about 20% in both AS graphs, and, furthermore, half of the countries have fewer than 10 ASes.

We depict the number of countries in each AS community in Figure 6(d). The communities with fewer than 100 ASes have ASes from fewer than 5 countries, while the number of countries rapidly grows in the communities of higher than 100 ASes. We consider two reasons for low locality in large communities. One is the high proportion of countries that have only a few ASes; 20% of countries have only one AS. These ASes always fall in a community with ASes from different countries, and the geographical locality decreases. The other is economically and politically motivated alignment between

countries, such as EU. The ASes in Europe densely connect to each other, and African ASes also densely connect to European ASes probably due to strong economic ties. In Europe AS 1299 (TeliaNet Global Network) and AS 3257 (TINET) are the two major ASes. They connect ASes of 48 and 56 countries, respectively. They both fall into R7 in Figure 5(e), which rightly reflects the important role these ASes play in regional Internet business sectors.

Meunier *et al.* report that intermediate communities produced in the greedy agglomerative process of the Louvain method can be useful to understand the structure of a network [26]. Following their lead, we investigate the sub-communities of the largest community in the AS graph. We observe that the sub-community sizes are just one-tenth of the largest community. Also the number of countries in a community decreases similarly. This indicates that Louvain method begins with a geographically concentrated community in early stage and gradually merges near communities step by step.

Class	Tier 1	Large ISP	Small ISP	Stub
R1	0	1	166	22,027
R2	0	6	292	2,197
R3	0	2	43	3,103
R4	0	72	693	3,628
R5	0	17	20	10
R6	0	5	15	2
R7	8	141	83	27

TABLE III
ASES (O) IN Z-P PLANE

Class	Tier 1	Large ISP	Small ISP	Stub
R1	0	5	326	17,844
R2	0	12	228	930
R3	0	4	39	3,165
R4	0	69	587	1,973
R5	0	19	12	4
R6	0	3	4	2
R7	8	128	70	18

TABLE IV
ASES (D) IN Z-P PLANE

In Section II-C we give an overview of z-P analysis of the twelve networks. Here we verify how a structural role of each AS defined in the z-P space matches the conventional classification. Oliveira *et al.* classify ASes into four classes based on the number of customers; a *tier 1*, a *large ISP*, a *small ISP* and a *stub*. We find some correlations between classes in z-P space and the classification by Oliveira. First, tier-1 ASes and large ISPs are mostly classified as $R7$. Nineteen out of top 20 ASes listed in the AS ranking page of CAIDA [38] are also in R7. A node in R7, by definition, connects well to other nodes in its own community and also to other communities. This matches the definitions of tier-1 by Oliveira and top ASes by CAIDA. Second, an AS of a small ISP is distributed in non-hub regions. This implies that most ASes of a small ISP are clustered with ASes of tier-1 or large ISPes, since the z-score is defined as the relative proportion of within-community degree. Finally, most stubs are classified as R1. They have few links to other ASes, and their z-scores and participation coefficients are low.

IV. Conclusion

In this paper we have investigated mesoscale structural characteristics of communities in networks. We have found that community size distributions do not have a uniform distribution; some (Orkut and Cyworld) follow power-law and other do not. However, the question about the origin of the power-law nature in community size remains. Using two AS graphs for community validation, we demonstrate that roles defined in the z-P analysis match closely those that ASes play regionally and offer more detailed view. We believe this work is just a first step towards much research into structural properties of network topologies and will continue our investigation in this direction.

Acknowledgment

This work was supported by NAP of Korea Research Council of Fundamental Science & Technology.

References

[1] H. Kwak, Y. Choi, Y.-H. Eom, H. Jeong, and S. Moon, "Mining communities in networks: a solution for consistency and its evaluation," in *IMC '09: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. New York, NY, USA: ACM, 2009, pp. 301–314.

[2] R. Guimera and L. A. Nunes Amaral, "Functional cartography of complex metabolic networks," *Nature*, vol. 433, pp. 895–900, 2005.

[3] R. V. Oliveira, B. Zhang, and L. Zhang, "Observing the evolution of internet as topology," in *SIGCOMM '07: Proceedings of the 2007 conference on Applications, technologies, architectures, and protocols for computer communications*. New York, NY, USA: ACM, 2007, pp. 313–324.

[4] A. Dhamdhere and C. Dovrolis, "Ten years in the evolution of the internet ecosystem," in *IMC '08: Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*. New York, NY, USA: ACM, 2008, pp. 183–196.

[5] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, no. 6, p. 066111, Dec 2004.

[6] K. Wakita and T. Tsurumi, "Finding community structure in mega-scale social networks," *CoRR*, vol. abs/cs/0702048, 2007.

[7] 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*, vol. 2008, 2008.

[8] S. Fortunato and M. Barthlemy, "Resolution limit in community detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 1, pp. 36–41, 2007.

[9] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, "On modularity clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 2, pp. 172–188, 2008.

[10] Y. Diao, M. Li, Z. Feng, J. Yin, and Y. Pan, "The community structure of human cellular signaling network," *Journal of Theoretical Biology*, vol. 247, no. 4, pp. 608–615, 2007.

[11] A. Chin and M. Chignell, "A social hypertext model for finding community in blogs," in *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*. New York, NY, USA: ACM, 2006, pp. 11–22.

[12] D. M. Chavis and A. Wandersman, "Sense of community in the urban environment: a catalyst for participation and community development," *American Journal of Community Psychology*, vol. 18, pp. 55–81, 2002.

[13] R. Guimera, S. Mossa, A. Turtschi, and L. A. N. Amaral, "The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 22, pp. 7794–7799, 2005.

[14] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002. [Online]. Available: http://www.pnas.org/content/99/12/7821.abstract

[15] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2658–2663, 2004.

[16] K. Yuta, N. Ono, and Y. Fujiwara, "A gap in the community-size distribution of a large-scale social networking site." [Online]. Available: arXiv:physics/0701168v2

[17] Y. Hu, H. Chen, P. Zhang, M. Li, Z. Di, and Y. Fan, "Comparative definition of community and corresponding identifying algorithm," *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, vol. 78, no. 2, p. 026121, 2008.

[18] I. F. Gergely Palla, Imre Dernyi2 and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814–818, 2005.

[19] S. L. Yong-Yeol Ahn, James P. Bagrow, "Link communities reveal multi-scale complexity in networks," 2009. [Online]. Available: arXiv:0903.3178v2

[20] A. Arenas, L. Danon, Authorz-Guilera, A. Gleiser, and A. Guimera, "Community analysis in social networks," *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 38, pp. 373–380, 2004.

[21] M. B. J.-C. Delvenne, S.N. Yaliraki, "Stability of graph communities across time scales," 2008. [Online]. Available: arXiv:physics/08121811v4

[22] A. C. Benjamin H. Good, Yves-Alexandre de Montjoye, "The performance of modularity maximization in practical contexts," 2009. [Online]. Available: arXiv:physics/09100165v1

[23] Y.-H. Eom, C. Jeon, H. Jeong, and B. Kahng, "Evolution of weighted scale-free networks in empirical data," *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, vol. 77, no. 5, p. 056105, 2008.

[24] K.-I. Goh, Y.-H. Eom, H. Jeong, B. Kahng, and D. Kim, "Structure and evolution of online social relationships: Heterogeneity in unrestricted discussions," *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, vol. 73, no. 6, p. 066123, 2006.

[25] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, pp. 452–473, 1977.

[26] D. Meunier, R. Lambiotte, A. Fornito, K. D. Ersche, and E. T. Bullmore, "Hierarchical modularity in human brain functional networks," *Frontiers in Neuroinformatics*, vol. 3, no. 37, 2009.

[27] S. Milgram, *Psychology Today*, vol. 1, no. 1, pp. 61–67, 1967.

[28] M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, no. 6, p. 1360, 1973.

[29] M. A. Rodriguez and A. Pepe, "On the relationship between the structural and socioacademic communities of a coauthorship network," *Journal of Informetrics*, vol. 2, no. 3, pp. 195 – 201, 2008.

[30] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabasi, "Hierarchical Organization of Modularity in Metabolic Networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[31] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 21, pp. 12 123–12 128, 2003.

[32] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*. New York, NY, USA: ACM, 1999, pp. 251–262.

[33] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, "The origin of power-laws in internet topologies revisited," in *IEEE Infocom*, 2002.

[34] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, "A model of Internet topology using k-shell decomposition," *Proceedings of the National Academy of Sciences*, vol. 104, no. 27, pp. 11 150–11 154, 2007.

[35] G. Siganos, S. Tauro, and M. Faloutsos, "Jellyfish: A conceptual model for the as internet topology," *Journal of Communications and Networks*, vol. 8, no. 3, pp. 339–350, 2006.

[36] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat, "Systematic topology analysis and generation using degree correlations," in *SIGCOMM '06: Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*. New York, NY, USA: ACM, 2006, pp. 135–146.

[37] D. Magoni and J. J. Pansiot, "Analysis of the autonomous system network topology," *SIGCOMM Comput. Commun. Rev.*, vol. 31, no. 3, pp. 26–37, 2001.

[38] "AS Ranking, Caida," http://as-rank.caida.org/.