1P-03

コミュニティ検出とその時間発展に基づくインターネットの成長過程の検証

濱本　拓海 †　　井上　寛康 ‡
兵庫県立大学大学院 シミュレーション学研究科 †‡

1　はじめに

インターネットは自律システム (Autonomous System:AS) によって構成され，そのトポロジーがどのように発展していくかは安定的な運用のために重要である．AS トポロジーに，トラフィックを考慮したコミュニティ検出をしたとき，コミュニティ内は高トラフィック，コミュニティ間は低トラフィックであることが期待される．このコミュニティが時間発展により形状が変わるならば，トラフィックの高低が変化することに対応し，それを予測することは重要となってくる．

インターネットのコミュニティ検出については先行研究がある [1] が，その時間発展の原因についてはまだあまり調べられていない．

2　データ

使用するデータはオレゴン大学の Route Views Archive Project[2] によって 2 時間おきに収集されている BGP (Border Gateway Protocol) データのスナップショットのうち，2009 年から 2018 年の各年 5 月 28 日 12 時時点のものである．BGP はネットワーク間の経路情報を交換するためのプロトコルであり，BGP データには各 AS から広告される経路情報が含まれる．BGP データを bgpdump[3] を用いて展開し，個々の AS をネットワークのノード，AS_PATH 属性で連続した AS をネットワークのリンクとして抽出する．

3　方法

3.1　コミュニティ抽出

ネットワークからコミュニティを抽出するために Infomap 法を用いる．Infomap はランダムウォーカーと情報理論を使用する．

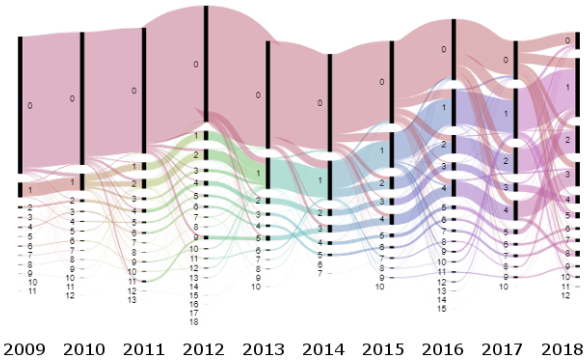

図 1　AS ネットワークのコミュニティ変遷

$$L(C) = q_{\curvearrowright} H(C) + \sum_{i=1}^{m} p^i_{\circlearrowleft} H(\mathcal{P}^i), \tag{1}$$

$L(C)$ はランダムウォーカーがネットワークのノード間のリンクを通ってコミュニティを移動するときの平均記述長を表す．右辺の第 1 項はコミュニティ間のランダムウォーカーの動きを表し，$q_{\curvearrowright}$ はランダムウォーカーがコミュニティを切り替える確率，$H(C)$ は与えられたコミュニティインデックスの平均記述長である．第 2 項はコミュニティ内のランダムウォーカーの動きを表し，$p^i_{\circlearrowleft}$ はコミュニティ C_i 内を動く割合，$H(\mathcal{P}^i)$ はモジュール i のエントロピーを示す．

リンクデータを各年それぞれ Infomap からコミュニティ抽出を行う．最も大まかなレベル 1 のコミュニティを，RAW[4] を用いて年ごとに追跡する Alluvial 図を作成すると図 1 のようになる．各年におけるコミュニティを数字で示し，0 はその時点では存在しないが，いずれかの時点では存在する AS の集合である．図ではコミュニティが分裂や吸収に伴ってコミュニティ数を変化させながら拡大している様子がわかる．ここからは 2009 年から 2017 年まで，それぞれの年と翌年の 2 時点に共通して存在するリンクを抽出する．

3.2　回帰と変数

あるコミュニティが，次の時点においてどれだけ分裂した状態となるかを，Lasso 回帰によって特徴選択およびモデルの推定を行う．被説明変数はコミュニティの

Growth process of the Internet based on community detection and its temporal evolution
†Takumi Hamamoto　‡Hiroyasu Inoue
†‡Graduate School of Simulation Studies, University of Hyogo

分裂度とし，それを示す指標としてハーフィンダール・ハーシュマン指数 (HHI) を用いる．HHI は 0 から 1 の範囲の値をとり，1 であるとき，あるコミュニティが次の時点においても分裂が起こらないことを示し，0 に近付くほど次の時点で細かく分裂している状態を示す．ある時点 t におけるコミュニティ C の HHI h は，C に属するノードが $t+1$ の時点で n 個のコミュニティに分裂し，分裂前のコミュニティ C のノード数のうち分裂後の複数のコミュニティに含まれるノード数の割合を s_c とおくと以下の式で示される．

$$h = \sum_{c=1}^{n} s_c^2 \tag{2}$$

次に説明変数を，コミュニティ境界リンク数 (fringe_node)，コミュニティ境界平均トラフィック (fringe_traffic)，コミュニティ境界リンクの媒介中心性平均 (fringe_betweenness)，コミュニティ境界ノード数 (fringe_node)，コミュニティ内リンク数 (internal_link)，コミュニティ内平均トラフィック (internal_traffic)，コミュニティ内リンクの媒介中心性平均 (internal_betweenness)，コミュニティノード数 (node)，コミュニティ境界リンクの割合 (fringe_link_ratio)，コミュニティ境界ノードの割合 (fringe_node_ratio)，コミュニティ内部リンクの割合 (internal_link_ratio)，別のコミュニティと接続されていないノードの割合 (internal_node_ratio)，内部リンクの平均次数 (fringe_link_mean)，境界リンクの平均次数 (internal_link_mean) とする．

AS 間のトラフィックは一般には公開されていないため，ここでは重力モデル [5] から計算したトラフィックを使用する．説明変数における各トラフィックは重力モデルにより以下の式で表される．

$$e_{ij} = \gamma \cdot x_i \cdot x_j \tag{3}$$

ここで，e_{ij}はトラフィック量，x_iおよび x_j はそれぞれ AS i，AS j の次数を表す．γ はスケーリングファクターであり，ここではすべて 1 とし，つまりトラフィックは単に次数の積で表される．

4 結果および考察

2009 年から 2017 年までの各年とその 1 年後のコミュニティ変化から計算した HHI について R の glmnet ライブラリを使用して Lasso 回帰を行う．データ数は 106 であり結果を表 1 に示す．

表 1 Lasso 回帰による変数選択と係数

変数	係数
fringe_link	.
fringe_traffic	3.359229e − 06
fringe_betweenness	9.551754e − 09
fringe_node	−8.892018e − 05
internal_link	−8.234010e − 06
internal_traffic	3.444343e − 05
internal_betweenness	1.458689e − 07
node	.
fringe_link_ratio	−2.217738e + 00
fringe_node_ratio	3.463105e − 01
internal_link_ratio	3.755002e − 02
internal_node_ratio	−3.892685e − 02
fringe_link_mean	7.894059e − 01
internal_link_mean	−2.699508e − 01

結果から境界リンク数とノード数は変数選択されなかった．正の係数である変数はその値が増加すると HHI が大きくなり分裂しにくい状態になりやすいことを示し，負の係数である変数はその逆である．コミュニティ境界リンクの割合が増え，内部リンクの割合 (internal_link_ratio) が減ると HHI は減少し，分裂しやすくなる．境界にてトラフィックや媒介中心性が増えることは分裂しやすさには直結しておらず，むしろリンクが内部で多いことがコミュニティの分裂にとって重要になってきており，境界と接触がないノードが多いことはそれ自体は分裂を抑えることがないことがわかる．つまり，分裂が起こる際には，境界部分の構造だけが影響しているのではなく，内部の構造についても影響しているということである．

今後の課題として，現在はコミュニティの分裂のみであるが，合流に関しても重要なので引き続き取り組む．

参考文献

[1] Yu Nakata, Shin’ichiArakawa, Masayuki Murata. Analyzing the evolution and the future of the internet topology focusing on flow hierarchy. *Journal of Computer Networks and Communications*, Vol. 2015, p. 2, 2015.

[2] University of oregon route views archive project. http://archive.routeviews.org/.

[3] RIPE NCC. bgpdump. https://bitbucket.org/ripencc/bgpdump/wiki/Home.

[4] DensityDesign Lab. Raw. http://app.rawgraphs.io/.

[5] Matthew Roughan, Albert Greenberg, Charles Kalmanek, Michael Rumsewicz, Jennifer Yates, and Yin Zhang. Experience in measuring internet backbone traffic variability: Models metrics, measurements and meaning. In *Teletraffic Science and Engineering*, Vol. 5, pp. 379–388. Elsevier, 2003.