論文の内容の要旨

論文題目 Stochastic processes on complex networks
（複雑ネットワーク上の確率過程）

氏　名　　川本　達郎

本研究では、複雑ネットワーク上でのダイナミクスやその構造を調べるための確率過程について議論する。二部構成となっており、第一部では、オンラインソーシャルネットワークを念頭に置いた、複雑ネットワーク上での情報拡散モデルの提案とその解析について述べる。第二部では、ネットワークのコミュニティ構造をランダムウォークを用いて検出する、map equation と呼ばれる手法について、種々の解析的な結果を述べる。

第一部の根底にあるのは Twitter データを始めとするビッグデータの存在である。よく知られている確率過程に対して、ベキ則に従う次数分布などの、複雑ネットワークで比較的一般に見られる性質がどのような影響を与えるかについては既に多くの研究がなされている。一方、近年ビッグデータと呼ばれる種類のデータの出現により、今までは見ることができなかった、ネットワーク上での詳細なプロセスの統計的解析が可能になった。これにより、複雑ネットワーク上の基本的なダイナミクスについて、よく知られている確率過程以上のものを観測することができるようになったのである。そこで、本研究では特にオンラインソーシャルネットワーク上での情報拡散に注目し、著者が Twitter から集めたデータを基に、古典的な分岐過程とは異なるタイプの拡散モデルを提案した。

本研究で考えるモデルは、種ノード (図 1 の円の中心のノード) から情報が発信され、それがネットワーク上のリンクを伝わって拡散していく様子を記述する、非常にシンプルな過程である。種ノードからの距離で周辺のノードを分類し、距離 g にいて、かつ情報を受け取ったノードの数を N_{g-1} とする。種ノードから情報が発信されると、まずリンクが繋がっている N_0 個の

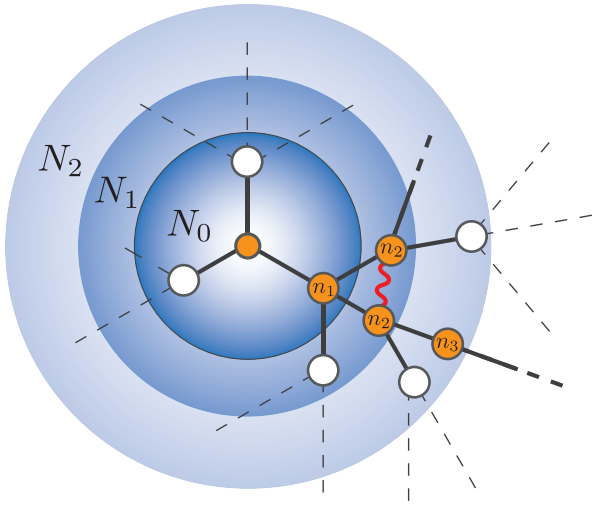

図 1: オンラインソーシャルネットワーク上の情報拡散モデルの概念図

ノードに伝わる。そのうちで n_1 個のノードが情報を他の繋がっているノードへと伝え、それによって新たに情報が伝わったノード数が N_1 である。その N_1 個のノードのうち、さらに情報を伝えるノードが n_2 個出現すると、さらに情報は拡散する。拡散ノードがいなくなるまでこれが繰り返される。ここで、独立な正の確率変数 β_g を導入し、距離 g に居る拡散ノード数 n_g が（ノード数 N_{g-1} で規格化して）

$$n_g = \beta_g N_{g-1} \tag{1}$$

として与えられるとすると、n_g は $\beta_1, \beta_2, \ldots, \beta_g$ を用いて、ランダム乗算過程として表すことができる。

提案モデルの特徴の一つは、固定した種ノードからの情報拡散を記述する、局所的なモデルになっていることである。これは種ノードをランダムに選ぶいくつかの先行研究のモデルとは対照的である。また、古典的な分岐過程は、各ノードのイベントは互いに独立であるとし、それぞれのノードに確率変数を割り当てた微視的なモデルである。しかし実データでは、それでは説明できない集団的な振る舞いが見られた。提案した確率モデルは、種ノードからの距離 g で世代分けされたノード集合に対して確率変数を割り当てており、そのような集団的振る舞いを記述することができる巨視的なモデルとなっている。

第一部の後半では、提案した確率モデルから導かれる数理的な性質を議論する。理想化された条件下においては、あるパラメター以上で平均的には拡散が止まらなくなる転移点が存在する。これはいわゆる「炎上」と呼ばれる現象に対応していると捉えることができる。確率変数

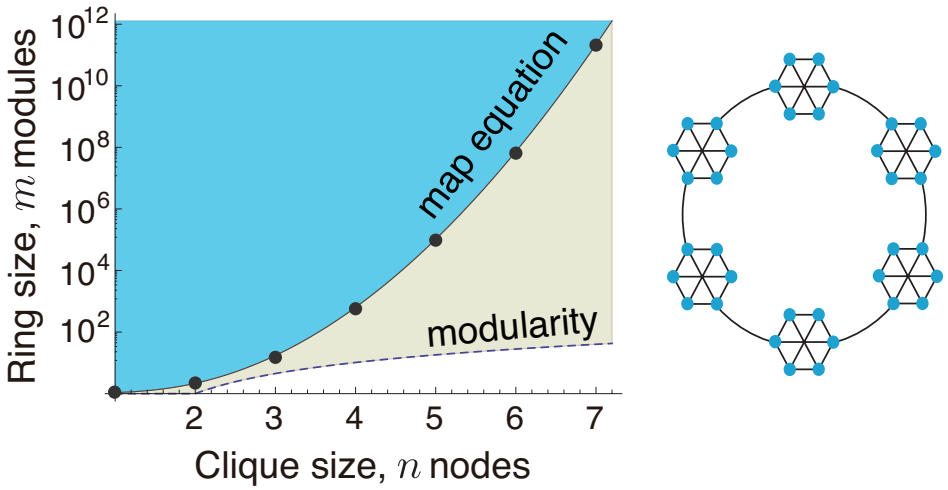

図 2: リングでの map equation と modularity の分解能限界

β_g が互いに独立である場合、ノードの平均次数を $\overline{k}$ すると、臨界的な β_g の期待値 $\beta_{\rm ex}$ は、

$$\beta_{\rm ex} = \overline{k}^{-1} \tag{2}$$

であるが、β_g が互いにある形の正の相関を持つ場合、

$$\begin{aligned}\beta_{\rm ex} &= \overline{k}^{-1}\left(\frac{\langle\beta_g\rangle\langle\beta_{g+1}\rangle}{\langle\beta_g\beta_{g+1}\rangle}\right)\\ &= \overline{k}^{-1}\left[1+\rho(\beta_g,\beta_{g+1})\frac{V(\beta_g)}{\langle\beta_g\rangle^2}\right]^{-1},\end{aligned} \tag{3}$$

となり、転移点は下がることが分かる。ここで $\langle\cdots\rangle$ は統計平均であり、$V(\cdots)$ は分散、$\rho(\beta_g,\beta_{g+1})$ は β_g と β_{g+1} の相関係数を表す。

第二部の対象である map equation は、コミュニティ検出手法の一つであり、近年のベンチマークによって非常に強力であることが示されている。Map equation はネットワーク上でのランダムウォークを考え、その符号長から情報理論的にコミュニティ分割の良さを評価し、その評価関数を最小化する分割が解となる。この最適化問題の振る舞いは、主に実験的に調べられ、手法が内包している理論的性質はあまり明らかにされていない。特に、分割数が自動的に定まるタイプのコミュニティ検出手法では、一般に分解能限界と呼ばれる特徴的なスケールを持つことが知られているが、今まで map equation がどのスケールに、どのような依存性で分解能限界を持つのかは知られていなかった。

第二部の主要な結果は、map equation の分解能限界の解析的な評価である。これにより、他の有名な評価関数である modularity がネットワーク全体のリンク数に依存した分解能限界を持

つのに対し、map equation の分解能限界はカットサイズと呼ばれる、コミュニティの間を繋ぐリンクの総数によって決まっていることが明らかになった。具体的には

$$\frac{2^{2l_c+\epsilon}}{l_c+1} \lesssim C \tag{4}$$

という形で表される。ここで、l_c はコミュニティ内部のリンク数、C はカットサイズ、$\epsilon \simeq 0.1146$ である。これは、この不等式を満たすような l_c を持つコミュニティを map equation は検出できないことを意味している。また、定量的にも map equation の方が modularity よりも小さいコミュニティを検出する（分解能が高い）ことが分かる（図 2 参照）。

さらに第二部後半では、その分解能限界は実質的に解消できることを示す。元々の map equation ではなく、それを拡張した（原論文の著者らによって提案されている）階層型 map equation を用いることで、分解能限界を取り除くことができるのである。階層型 map equation は、元々の map equation が、ノード集合としてのコミュニティとネットワーク全体という二層構造を考えているのに対し、コミュニティのコミュニティ（supermodule）のような構造を導入したものである。この場合の分解能限界は、ネットワーク全体のカットサイズではなく、コミュニティを内包している supermodule の中だけのカットサイズに依存しており、実質的に分解能限界の問題は解消される。これらの理論的な結果は、公開されているプログラムコードを用いて求めた、実際のネットワークデータのコミュニティサイズ分布の様子からも確かめられる。