Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations (2005)

Jure Leskovec, Jon Kleinberg, and Christos Faloutsos.
Kleinberg @ Cornell Univ.

KDD2005 (http://www.cs.cornell.edu/ lars/kdd06-comm.pdf)

1 本論文の目的

時間軸を持った巨大なネットワークデータを分析し、ネットワークの成長に関する知見を得る。さらに知見から導き出された新しい法則性に即したネットワーク成長モデルを提案する。

2 背景

様々な研究により静的なグラフに関するパターンの発見がなされているが、多くは巨大なネットワークのあるスナップショットに対してのみである。しかし長期間におけるネットワークの成長に関する情報が無くては、これらの知見を現在のネットワークに適用するのは難しい。

そこで本研究では 4 つの実際に存在するネットワークの長期間にわたるスナップショットの分析を通してネットワークの成長に関する新しい法則性を見つける。このような知見は、グラフの生成・サンプリング・将来予測・異常検知などに役立つと考えられる。

3 実験 1: 実データの分析

3.1 手法・アルゴリズム

分析対象となるデータは (a) 論文アーカイブ arXiv の引用関係（1993 年 1 月から 2003 年 4 月まで。ノードは論文でエッジは論文間の引用関係。ただしデータセットに含まれない論文との引用関係は無視する。ノード数は 29,555、エッジ数は 352,807）、(b)National Bureau of Economic Research が管理している特許データの引用関係（1963 年 1 月から 1999 年 12 月まで。ノードは特許、エッジは特許間の引用関係。ノードは 3,923,922、エッジは 1975〜1999 の全員引用関係で 16,522,438）、(c)BGP のログから得た AS 間の通信関係（1997 年 11 月から 2000 年 1 月まで。ノードは AS で、エッジは AS 間の接続関係。引用グラフと異なり時が経てばノードやエッジが削除されることもある。）、(d)arXiv の 5 分野の論文の共著関係（1992 年 4 月から 2002 年 3 月まで。ノードは論文と著者の 2 種類、エッジは著作関係。分野は ASTRO-PH、HEP-TH、HEP-PH、COND-MAT、GR-QC の 5 分野で最小が GR-QC の 19,393 ノードの 26,169 エッジ、最大が ASTRO-PH の 57,381 ノードの 133,170 エッジ。）、の 4 種類である。

3.2 知見

データの平均出次数を調べたところ、年々増加しているという傾向が 4 つ全てに見られた (Fig.1)。また、ノード数とエッジ数の関係は冪則に従う傾向が見られた (Fig.2)。このとき、指数が 1.0〜2.0 になることを Densitification Power Low と呼ぶ。

ネットワークの直径を調べたところ、年々減少し、ある程度で収束傾向にあるのが 4 つ全てにみられた (Fig.3)。これだけでは直径の計算方法 (ASF)、小さなコンポーネント群の存在、全ての過去データがあるわけではない (missing-past)、という問題がある。そこで最大コンポーネントの直径を厳密に計測するという方法をとり、また、ある時間 t0 以前のデータがある場合と無い場合との比較を行った。結果、t0 以前のデータの有無にかかわらず減少し収束するという傾向が得られた。

4 実験 2: モデルの提案

4.1 手法・アルゴリズム

データ分析から得られた Densification Power Low と直径の収束という特徴を持つネットワーク成長モデルとして Forest Fire Model を新たに提案する。このモデルでは、ノード n が新たに追加されたとき、まずランダムでノード w を選択して out-link を追加する。次にノード w が持つエッジからランダムで x 本を選択し、そのエッジのもう一方の端のノードに対して out-link を追加する。追加したノードに対してはさらに同様の処理を行い、これを繰り返す。ただしその際、同じノードは 2 度と選ばれないようにする。これは引用論文が引用している論文を自分も引用する、という振る舞いを模したものといえる。

なお、選択するエッジ数 x は二項分布に従い、その生起確率 p を forward burning probability と呼ぶ。そして選んだエッジの中で in-link が選ばれる割合 r を backword burining probability と呼ぶ。提案モデルが持つパラメータはこの二つだけである。

4.2 評価

p と r のパラメータを変えることで、ノード数とエッジ数 (Fig.5 左)、ノード数と直径 (Fig.5 右) がそれぞれどのような関係になるかを調べた。結果、p を中程度 (p=0.37, r=0.32) にすることで、ノード数とエッジ数の間の冪則と直

径の収束が見られた。また、既存研究の知見が示すような、次数分布における冪則も見られた。

5　まとめ

本論文は大きく二つに分かれる。前半では 4 つのグラフデータを分析し、グラフの成長に関する新しい法則（グラフの緻密化、直径の収束）を見つけ出した。後半では見つけた法則に沿った新しいグラフモデルを提案し、提案モデルがその法則性を持っていることを示した。

（文責：濱崎雅弘）