共引用ネットワークを用いたSleeping Beauty論文の解析

小舘 俊[1], 木下 賢吾[1,2]

1 東北大学 大学院情報科学研究科　2 東北大学 東北メディカル・メガバンク機構

学術論文におけるSleeping Beauty (SB)とは、出版されてからしばらくの間はあまり引用されないが、ある時から急激に引用されるようになる論文を指す。先行研究によりこれらSB論文はごく少数の例外ではないことが示唆されているが、その発生のメカニズムは明らかではない。今回我々はSB同士の関連を調べるため、共引用ネットワークを用いて解析を行なった。各種指標により、SB論文から成る共引用ネットワークは通常の論文のものと比べて強い結びつきをもっていることが分かった。解析結果から、SB論文が多く存在していても、それらの現象を駆動しているのは比較的少数のイベントである可能性が示唆された。これらの知見は、科学発展の歴史や過程のより包括的な理解につながると考えられる。

はじめに：論文の引用関係とSleeping Beauty

学術論文は科学者たちの研究成果の結晶である。ある論文に記載された成果はその著者だけによって生み出されたわけではなく、先行研究という歴史の上に成り立っており、それらは引用という形で論文にあらわれている。このようなデータを集積し解析することで、科学の発展について俯瞰的な知見を得ることができる[1,2]。たとえば各論文の被引用数を分布として見ると、いわゆるべき乗則に従い少数の論文が多数の引用を受けている。これは多く引用された論文はそれだけ人目に付きやすくなり更に引用されやすくなる、つまり「富めるものがさらに富む」という自然で普遍的なプロセスの結果として記述できる[3,4]。またこのプロセスに従う限りは、多数の引用を獲得できるのは出版直後に引用され始めた論文のみであることが知られており[5]、ある一論文の典型的な被引用履歴としては、出版直後にピークを迎えた後、新規性の低下により減少していくという過程をたどることになる[6]。

しかし被引用数が多い論文の中にも、出版直後にはあまり引用されず、ある時から急激に引用されるようになる、いわゆる"Sleeping Beauty (SB)"と呼ばれるものが存在する[7-10]。上記プロセスではこの現象は説明できないため、SBの挙動を解析することで科学の発展についてより包括的な理解を得られることが期待できる。このような理由から、SBについての調査を行なった。

近年、論文の"SBらしさ"（スコアB）を、パラメータフリーで算出する手法がKeらにより考案された（[10], 図1）。

$$B = \sum_{t=0}^{t_m} \frac{l_t - c_t}{\max\{1, c_t\}} \qquad (1)$$

本研究では、生物医学分野論文のデータベースであるMEDLINEのデータを対象に、この手法を用いて検出されたSBについて考察した。

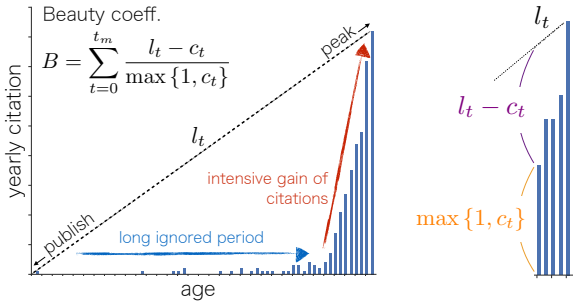

図1. SBの検出方法[8]。ある論文について年ごとの被引用数を用いて、出版年からピークに達する年までの値でスコアの計算を行なう。

Bの分布からは、SB論文と「通常の」論文を区別するような明確な境界線は見出だせなかった（図2）。つまり、SBはごく少数の例外ではなく、連続的に存在していることが示唆された。この結果は物理分野および科学全般の論文を対象とした先行研究と同様であるが、なぜこのような分布となるかは明らかになっていない。我々はこの原因として、あるSBの「目覚め」が別のSBの目覚め

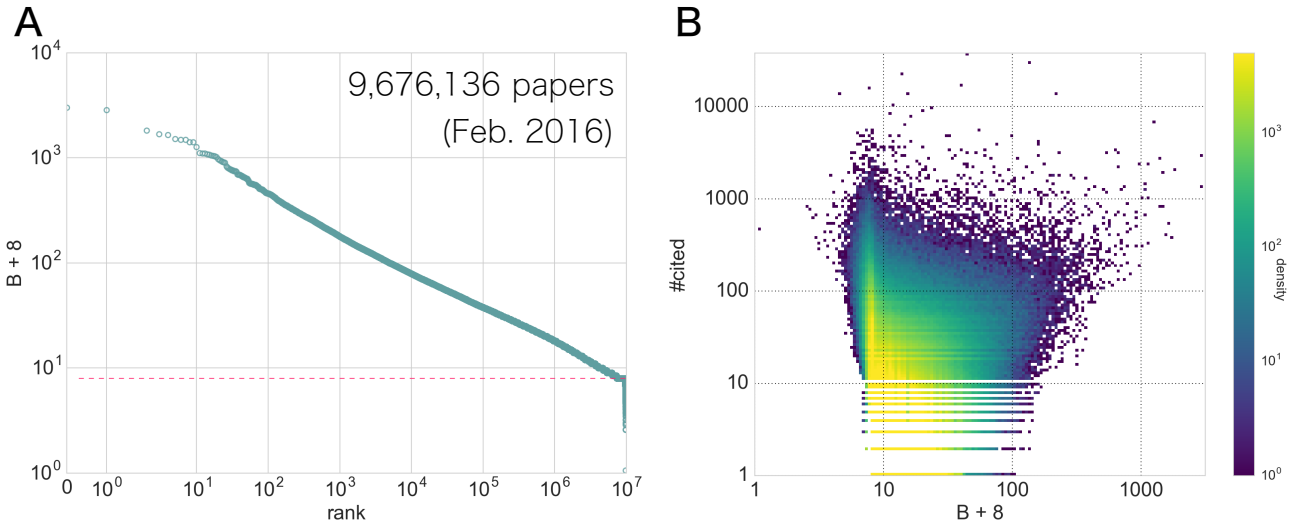

図2. (A) Bの分布, (B) Bと被引用数の分布。データセットは生物医学分野文献のデータベースであるMEDLINEから入手した。どちらの分布からも、SBを分離する明確な境界線は見出だせなかった。

を引き起こす、つまり、複数のSBが同時に引用されるようになることがあるためではないかと考えた。そこで、共引用という関係性を用いて、SBの挙動について解析を行なった。

Sleeping Beautyの共引用ネットワーク

ある論文2つがひとつの論文内で引用されているとき、引用されている2つの論文の間に成り立つ関係が共引用であり、論文同士の関連性を示す指標の一つである[11]。ネットワークとして考えるとエッジに重み（＝共引用の回数）が付いた無向ネットワークとなり、論文群が同時に引用されやすければ、ネットワークは結びつきが強いものとなる。

Bの分布からSBが連続的に存在していることは先に述べたが、ここではこの現象を駆動している、より影響力の強いと思われるSBを解析することを考え、スコア上位100本のSBを用いて共引用ネットワークを作成した。また比較対象として、同程度引用されている100本の論文をランダムに選び、それらから共引用ネットワークを作成した。これらを対象に、結びつきの強さを見るため、エッジの重みがしきい値以下のものを切断していき残ったネットワークの各種指標を計測した。

その結果、SBの共引用ネットワークは、ランダムな論文セットのものと比較して、強い結びつきをもっていることが分かった（図3）。これより、SBは同時に引用されることが多く、

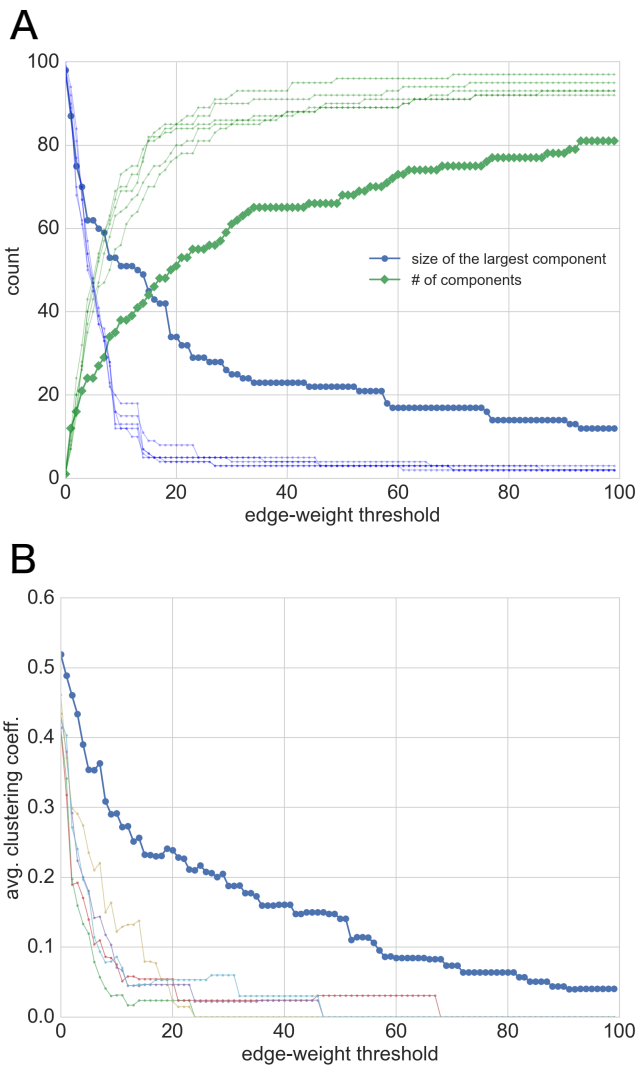

図3. SBの共引用ネットワークの結びつき。エッジの重みをしきい値として切断していったときの、(A) 最大コンポーネントのノード数(青)とコンポーネント数(緑), (B) 平均クラスタリング係数。薄い線は同程度引用されているランダムセット5つの値。

SB同士には強い関連があることが示唆された。またここからSBの目覚めを駆動しているのは論文の数と比較して少数のイベントであることが考えられるため、その解明を試みた。

目覚めの一例：知能・精神に関する検査について

上記SBの共引用ネットワークにおいてとりわけ強固な結びつきをもつ論文群について調査するため、重みが100以上のエッジのみ、すなわち100回以上共引用されているという関係のみを残したネットワークを作成した（図4）。すると多数のSBが接続された大きなコンポーネントが見られたため、これに含まれるSBについて詳しく調べることとした。

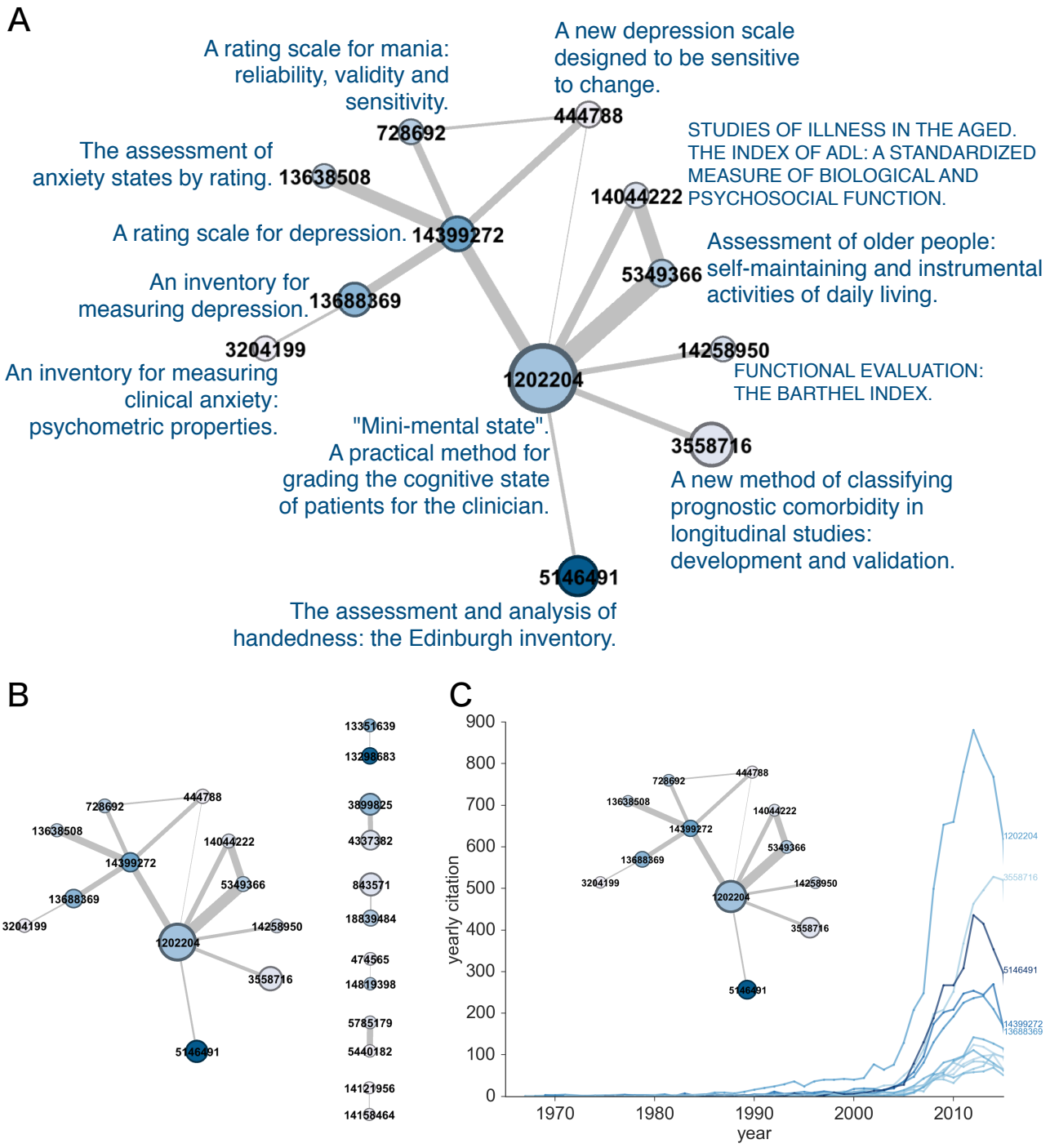

図4. SB上位100本での、100回以上の共引用をもつネットワーク。ノードの大きさは被引用数、色の濃さはBスコア、エッジの幅は共引用の回数に比例している。数字は論文のID (PubMed ID) 。(A) 最大コンポーネントの情報。(B) ネットワークの全体像。 (C) 最大コンポーネントに含まれる論文の被引用履歴。

まず、SBの目覚めを引き起こした存在として、後に出版されSBと一緒に引用されている論文が候補になると考えた。また仮説として、SBは異なる分野に「発見」されることで目覚めるというものがあり、実際にその傾向が確かめられている[10]。これらの考え方から、最大コンポーネントに含まれるSBの目覚めについて解析を行なった。

論文の分野が異なるという程度を定義するのに、ここではMeSH (Medical Subject Headings) を利用した。MeSHとは生物医学分野における用語集であり、MEDLINEに含まれる論文一つ一つに対し複数のMeSHの用語が付与されその内容を表している。このMeSHから、以下で定義されるJaccard類似度を計算することで、分野の類似度を測ることとした。

$$J(A,\ B) = \frac{|A \cap B|}{|A \cup B|} \qquad (2)$$

ここでAはあるひとつの論文に付与されたMeSHの集合、Bは上記最大コンポーネントに含まれるSB論文に付与されたMeSH全ての集合とした。これにより、最大コンポーネントのSB群と類似度の小さい、すなわち異なる分野の論文を検出した。

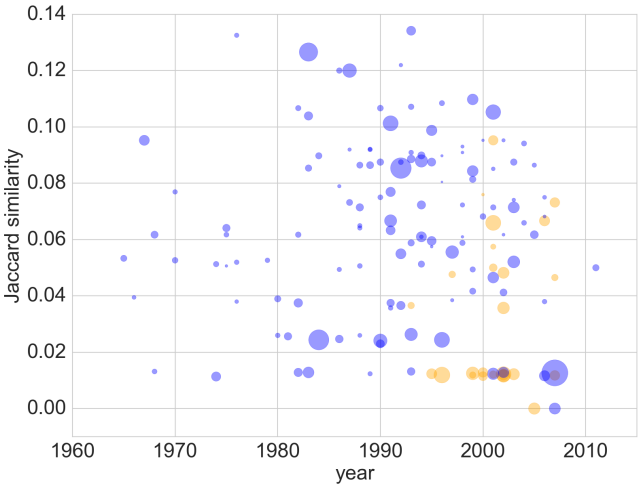

図5. 図4AのSB群と共引用されている論文の、出版年とSB群に対するJaccard類似度。SB群との共引用の回数が10回以上かつ次数が3以上のもののみを示した。またReview, Comparative study, SB上位100本に含まれる論文は除外した。オレンジ色の点は"Brain"または"Magnetic Resonance Imaging"の語を含むMeSHが付与された論文。点の大きさは被引用数に比例。Jaccard類似度の範囲が狭いように思えるが、ある論文に付与されるMeSHの数（=|A|）が十数個であるのに対し、SB群に付与されたMeSH全ての数(=|B|)が74個と多いためである。

図4Cの被引用履歴からは、2000年から2004年付近に出版された論文が目覚めを引き起こしたことがうかがえる。そこでこの期間に出版された、Jaccard類似度が小さい論文群に着目すると、そこには"Brain"や"Magnetic Resonance Imaging"といった語を含むMeSHが付与された論文が多く含まれていた（図5）。これより、最大コンポーネントに含まれるSB群–知能や精神に関する検査–は、脳の研究やMRI（特にfMRI）技術の発達に伴って、再び利用されるようになった、あるいは検査の対象である障害などの研究が発展した、という流れが見えた。

結論

本研究では、SBは共引用ネットワークにおいて強い結びつきをもっていること、またそれらSBの目覚めが比較的少数のイベントによって駆動されている一例を示した。したがって、SBという現象を解析する際には、SB論文・目覚めのトリガーとなった論文ともに、単一の論文ではなく複数論文のコミュニティとして解析することが必要である可能性が示唆された。

ただし今回解析した目覚めの例はあくまで一例であり、SBの全体像を把握するには至っていない。SBの目覚め方を分類するなどして現象への俯瞰的な理解を深めることが今後の課題である。

[1] Evans JA, Foster JG (2011) Metaknowledge. Science 331(6018):721–5.

[2] Zeng A, et al. (2017) The science of science: From the perspective of complex systems. Phys Rep 714–715:1–73.

[3] de Solla Price DJ (1965) Networks of Scientific Papers. Science (80-) 149(3683):510–515.

[4] Barabási A-L, Albert R (1999) Emergence of Scaling in Random Networks. Science (80-) 286(October):509–512.

[5] Newman MEJ (2009) The first-mover advantage in scientific publication. EPL (Europhysics Lett) 86(June):68001.

[6] Wang D, Song C, Barabási A-L (2013) Quantifying long-term scientific impact. Science 342(6154):127–32.

[7] Barber B (1961) Resistance by scientists to scientific discovery. Science 134:596–602.

[8] Cole S (1970) Professional standing and the reception of scientific discoveries. Am J Sociol 76(2):286–306.

[9] van Raan AFJ (2004) Sleeping Beauties in science. Scientometrics 59(3):467–472.

[10] Ke Q, Ferrara E, Radicchi F, Flammini A (2015) Defining and identifying Sleeping Beauties in science. Proc Natl Acad Sci USA 2015(35):40.

[11] Small H (1973) Co-citation in the scientific literature: A new measure of the relationship between two documents. J Am Soc Inf Sci 24(4):265–269.