

博士学位論文

局所構造に基づくネットワーク特徴による
実世界データの解析

東北大学大学院情報科学研究科

応用情報科学専攻

生命システム情報科学分野

小舘 俊

2020年1月20日

目次

要旨	4
1 導入	5
2 売買関係ネットワークの解析	9
本章の要旨	10
2.1 導入	11
2.2 材料と手法	13
2.2.1 データ	13
2.2.2 ネットワーク解析	13
2.2.3 ランダムフォレストによる分類	16
2.3 結果	18
2.3.1 記述統計	18
2.3.2 ユーザーの分類	27
2.4 考察	36
3 引用関係ネットワークの解析	39
本章の要旨	40
3.1 導入	41
3.2 材料と手法	43
3.2.1 データ	43
3.2.2 文献サブセットの構築	43
3.2.3 ネットワーク解析	45
3.3 結果	47
3.3.1 文献セットの基本統計量	47
3.3.2 引用関係ネットワークの構造	49
3.3.3 共引用関係ネットワークの構造	52
3.4 考察	59
4 総括	61
謝辞	64
参考文献	65
付録	73

目次

1	売買関係：有向三角形のパターンとその数え方	15
2	売買関係：各種ユーザーの次数	20
3	売買関係：各種ユーザーの隣人一人あたりの平均取引数	21
4	売買関係：各種ユーザーの sell probability	23
5	売買関係：各種ユーザーのローカルクラスタリング係数	24
6	売買関係：各種ユーザーの triangle congregation	25
7	売買関係：各種ユーザーの cycle probability	26
8	売買関係：通常ユーザーと不正ユーザーとを分類したときの ROC および PR 曲線	28
9	売買関係：分類器に用いた特徴量の permutation importance	30
10	売買関係：隣人間のつながり方を使わずに分類したときの ROC および PR 曲線	32
11	売買関係： $k_i = 1$ のユーザーを使わずに分類したときの ROC および PR 曲線	34
12	売買関係： $k_i = 1$ のユーザーと隣人間のつながり方を使わずに分類したときの ROC および PR 曲線	35
13	引用関係：文献セットの基本統計量	48
14	引用関係：引用関係ネットワークでの次数分布	49
15	引用関係：引用関係ネットワークでの連結成分に関する各種統計	51
16	引用関係：共引用関係ネットワークでの次数に関する分布	53
17	引用関係：共引用関係ネットワークにおけるコミュニティ分割の指標	54
18	引用関係：DR 文献セットの共引用関係ネットワークにおけるコミュニティの例	56
19	引用関係：DR 文献セットの共引用関係ネットワークにおけるコミュニティごとの被引用履歴	58

表目次

1	売買関係：各種ユーザーの統計情報	19
2	売買関係：分類器の AUC	29
3	売買関係： $k_i = 1$ のユーザーを除いた分類器の AUC	33
4	引用関係：共引用関係ネットワークにおける最大連結成分の頂点数	52
5	引用関係：DR 文献の共引用関係ネットワークにおけるあるコミュニティを構成する文献	55

要旨

インターネットや電子メール、ソーシャル・ネットワーキング・サービス (SNS)、各種データベース、計測機器などの普及・発達により、人間の様々な行動についてのデータが日々生み出され、利用可能になっている。これらを用いて、従来手法・環境では不可能だった、実世界に関する大規模なデータの解析を行ない、新たな知見を得ることが可能になってきた。このようなデータの多くは、本質的にネットワーク、すなわち人や物事の「つながり」として捉えることができる。ネットワーク（グラフとも呼ばれる）は、「つながり」を抽象化した、事物の関係性を記述できる柔軟な枠組みである。ネットワークに関する研究の潮流のひとつとして、現実に見られる性質を持つネットワークを生成するモデルをつくることで、最小の原理での現実世界の説明を目指すことがある。その一方で、ネットワークをデータ解析のための道具として用いる研究も盛んに行なわれている。たとえば、ネットワークを構成する頂点を特徴づける指標の計算を行ない、どの頂点が重要であるかなどの性質を調べることができる。本研究では、ネットワークを解析のための道具として用い、2つの実世界大規模データ：市場での売買関係と文献の引用関係を対象として、ネットワークの局所構造を用いた解析を行なった。売買関係ネットワークの解析では、オンラインの商取引市場であるメルカリのデータを解析した。数百の不正ユーザーと通常ユーザーのサンプルから、それぞれのユーザーの周りの、局所的なネットワークを構築した。これよりネットワーク上の指標を計算し、またそこから作成した特徴量を用いて機械学習により通常ユーザーと不正ユーザーの分類を行なった。引用関係ネットワークの解析では、学術論文のなかでも特異な性質をもつ delayed recognition (DR) 文献に着目した。DR 文献のセットを用いて、それらの間の引用関係ネットワーク、および共引用関係ネットワークを構築した。またランダム文献セットからも同様にネットワークを構築し、指標の比較を行なった。これらにより、制限がある中でネットワークを解析するための手法を提案し、その有効性を示した。

1 導入

インターネットや電子メール、ソーシャル・ネットワーキング・サービス (SNS)、各種データベース、計測機器などの普及・発達により、人間の様々な行動についてのデータが日々生み出され、利用可能になっている。これらのデータを用いて、従来手法・環境では不可能だった、実世界に関する大規模なデータの解析を行ない、新たな知見を得ることが可能になってきた。このような分野は計算社会科学と呼ばれ、活発な研究が行なわれている [1]。研究の一例を挙げると、2004 年のアメリカ大統領選挙期間中のブログの関係を解析し、保守派のブログは同じ保守派のブログへのリンクが、リベラル派のブログは同じリベラル派のブログへのリンクが多く、派閥間のリンクは比較的少数であること、すなわち派閥間の分断が見られることを示したものがある [2]。また科学そのものについても、文献などのデータが充実してきたことによって、引用関係や共著関係などから社会性を考えることが可能になってきた。このような分野は science of science と呼ばれ、近年発展してきている [3]。

上に挙げたデータの多くは、インターネットや SNS ではそれらの名前が示す通り、本質的にネットワーク、すなわち人や物事の「つながり」として捉えることができる。ネットワーク (グラフとも呼ばれる) は、「つながり」を抽象化した、事物の関係を記述できる柔軟な枠組みである [4]。数学の世界では古くから「グラフ理論」として研究が行われてきたが、現実の、しばしば大規模かつ複雑なネットワークを対象とした「ネットワーク科学」が発展してきたのは、比較的近年であると考えられる。発展のきっかけとして、現実世界で見られる様々なネットワークに共通する特徴を再現できる簡便なモデルが開発されたことが挙げられる。Watts と Strogatz は 1998 年、スモールワールド性、すなわちある頂点に隣接する頂点同士もまた隣接していることが多く、かつ任意の 2 頂点間の距離も比較的小さいという性質を再現するモデルを発表した [5]。Barabási と Albert は 1999 年、スケールフリー性、すなわち次数の分布がべき乗則に従うという性質を再現するモデルを発表した [6] (ただし、等価なモデルが先に Price によって発表されていた [7])。このように、ネットワークに関する研究の潮流のひとつとして、現実に見られる性質を持つネットワークを生成するモデルをつくることで、最小の原理での現実世界の説明を目指すことがある。

その一方で、ネットワークをデータ解析のための道具として用いる研究も盛んに行なわれている。たとえば、ネットワークを構成する頂点を特徴づける指標の計算を行ない、どの頂点が重要であるかなどの性質を調べることができる。著名な例として、Google のインターネット検索エンジンで使用されている PageRank がある [8, 9]。これはウェブページを頂点、ハイパーリンクを (有向) 枝としたネットワーク上のランダムウォークから計算される、頂点の重要度を表す指標である。また、ネットワークの中で密につながった部分構造を検出することで、データがどのようなまとまりに分けられるかを調べることができる。この

ような部分構造はコミュニティ構造 [10, 11] と呼ばれ、ネットワークの可視化と併せて関係性の直観的な把握にも便利である。このようにネットワーク解析は、データの構成要素の重要度を測ることや、それらの関係性を様々な側面から検証することによって、抽象的な数学の世界での研究のみならず、社会学 [12] や生物学 [13]、脳科学 [14] など、現実世界を扱う多様な分野に応用されている。

ネットワーク（の頂点）を特徴づける指標を大別すると、大域的なものと局所的なものに分けられる。大域的な指標とは、計算にネットワーク全体の構造を必要とするものである。例としては、上に記した PageRank やコミュニティ構造のほか、2 頂点を結ぶ最小の枝数の平均である平均距離、ネットワーク全体の三角形の多さを示すクラスタリング係数、他の頂点への距離の近さを表す近接中心性、二頂点を結ぶ最短経路にどれくらい寄与するかを示す媒介中心性などが挙げられる。局所的な指標とは、ある頂点あるいはその周辺の構造だけを用いて計算されるものである。例としては、その頂点をもつ枝の数である次数、その頂点をもつ三角形の多さを示すローカルクラスタリング係数などがある。また、ネットワークを構成する頂点および枝の一部に着目することも考えられる。

本研究では、ネットワークを解析のための道具として用いる立場をとる。加えて本研究では、ネットワークの局所的な構造および特徴に着目する。ネットワークを解析するにあたっては、大域的な指標と局所的な指標とをあわせて考えることで、より細かい考察ができることが基本的には考えられる。局所的な構造・特徴に着目する理由として、ここでは2つが挙げられる。ひとつは、利用できるデータが限られる場合という実用的な理由である。これは、ある空間を構成するデータがあるが、計算コストなどの都合でそのすべてを利用することはできない場合、部分的な情報のみから計算できる量を用いるということである。もうひとつは、全データ中の特定のデータのみに着目したい場合という、探究的な理由である。すなわち全データのうち、ある特定の興味ある性質を示す存在を対象として新たな知見を得たい場合、それらの間あるいはそれらの周辺の情報を特に用いるということが考えられる。

以上の動機から、本研究では2つの実世界大規模データ：市場での売買関係と文献の引用関係を対象として、ネットワークの局所構造を用いた解析を行なう。売買関係ネットワークの解析では、オンラインの商取引市場であるメルカリのデータを解析する。メルカリは日本最大の規模をもつオンライン消費者間取引市場である。この市場ではユーザーは互いに直接、多様なアイテムを売買することができる。しかし中には出品を禁止されている、不正なアイテムを扱うユーザーも存在し、その検出は管理者にとって重要な課題である。本研究ではユーザー同士の取引関係からつくられるネットワークを解析し、不正なアイテムを取引したユーザーの、ネットワーク上での特徴を探る。またそこで得られた特徴から、機械学習

の手法も用いて不正ユーザーの検出を試みる。引用関係ネットワークの解析では、学術論文のなかでも特異な性質をもつ delayed recognition (DR) 文献に着目する。これは、出版直後にはほとんど引用されないが、ある時から急激に引用されるようになった文献を指す。このような DR 文献はごく少数の例外ではなく、多数存在することが示唆されている。しかし既存の研究は、個別の DR 文献が着目されるきっかけとなった文献の検出が中心であり、DR 文献の性質や関係性は不明な面が未だ多い。本研究では、DR 文献を頂点とした引用関係ネットワークおよび共引用関係ネットワークの解析を行なう。ランダムな文献を頂点としたネットワークとの比較により、DR 文献に特有の性質を考察する。

2 売買関係ネットワークの解析^{*1}

^{*1} 本章の研究は、著者がデータ科学国際共同大学院プログラムにより University of Bristol に留学した際に行われた。計算資源として、University of Bristol の Advanced Computing Research Centre のものを用いた。本章の内容は以下の文献に基づく：
Kodate S, Chiba R, Kimura S, & Masuda N.
Detecting problematic transactions in a consumer-to-consumer e-commerce network. (submitted)

本章の要旨

オンライン市場の管理者は、不正なアイテムや存在しないアイテムを出品するなど、問題ある取引の一掃に日々取り組んでいる。不正行為を検出するための典型的な方法は、登録されたユーザーのプロフィールや、ユーザーの行動、各出品に付されたテキストや画像などを解析することであった。しかし、悪意あるユーザーはこれらの情報を隠したり偽装したりすることがあり得るため、この慣習的な方法には限界があることが考えられる。このような背景から、ネットワークを利用した指標が様々なオンライン取引プラットフォームで不正検出のために利用されてきた。本研究で我々は、売り手と買い手が有向枝で結ばれる、オンライン消費者間取引市場のユーザーがつくるネットワークの解析を行なった。我々は数百の不正ユーザーと通常ユーザーのサンプルから、それぞれの局所的な、いわゆる egocentric network を構築した。これより、中心頂点の隣接頂点間のつながり方までを用いる、8個のネットワーク上の指標を計算した。これらのネットワーク指標の記述統計に基づき、12個の特徴量を用いてランダムフォレストにより分類器を作成し、通常ユーザーと不正ユーザー（すなわち、4種類の不正アイテムのいずれかに関わったユーザー）の分類を試みた。分類器は不正の種類によらず、高い確度で通常ユーザーと不正ユーザーを分類することができた。

2.1 導入

オンラインでの商取引やコミュニケーションの急速な拡大に伴って、その不正な利用もまた急激に増加し、人々の生活に入り込んでいる。サイバー犯罪を含む不正行為は、年間数十億ドルもの損失を発生させ、社会を脅かしている [15,16]。特に、近年のオンラインでの活動が強い影響をもつ時代にあって、システムを攻撃することにはそれほど高いコストがかからないが、防御することには高いコストがかかる [17]。不正行為の範囲は広く、クレジットカードの不正利用、マネーロンダリング、コンピュータへの不正侵入、剽窃など多岐にわたる。

不正を検出・予防するために、計算機的・統計的な手法が長年開発されてきた [18–21]。不正検出のための標準的な方法は、機械学習を含む統計的な手法を利用することである。特に、不正なサンプルと不正でないサンプルの両方が利用可能な場合、教師あり学習によって分類器を作成することができる [18–21]。このような統計的分類器に与える特徴量の例として、クレジットカードでは取引額や、曜日、商品の種類、ユーザーの住所など、電話では通話回数や、通話時間、通話の種類、ユーザーの年齢、性別、地域など、オンラインオークションではユーザーのプロフィールや、取引履歴などが挙げられる [20]。

しかし、このような特徴量の多くは、高度な不正ユーザーには簡単に偽装されうる [22,23]。さらに、不正な行為を示す特定の語句の検出を管理者や当局が試みても、不正ユーザーはそれから逃れることに長けている [24–26]。たとえば、当局が薬物を意味する用語を見つけても、不正ユーザーは攪乱のため、簡単に別の用語に切り替えることができる。

ネットワーク解析は特徴量を構築する別の方法であり、不正検出の技術として新しいものではない [22,27]。考え方としては、データにおける頂点（ユーザーや商品など）の間のつながり方を使い、グラフ理論的な量を計算することで、それら頂点を特徴づけるというものである。これらの手法は、不正に関わる頂点が通常の頂点とは異なるつながり方のパターンを示すという期待によっている [22]。ネットワーク解析による不正検出が実践された分野の例を挙げると、保険 [28]、マネーロンダリング [29–31]、医療データ [32]、配車サービス [33]、社会保障システム [34]、モバイル広告 [35]、携帯電話ネットワーク [36]、SNS [37–40]、オンラインレビュー [41–43]、オンラインオークションおよび市場 [44–48]、クレジットカード取引 [49,50]、暗号通貨取引 [51]、その他様々な事例 [52] がある。一例として、あるオンラインオークションでは、不正ユーザーとその協力者が評価を上昇させるため、ユーザーのネットワークにおいて二部グラフに近い構造を形成することが示された [44]。次いで著者らは、このような疑わしいつながり方のパターンを検出するため、確率伝播法に基づくアル

ゴリズムを考案した。この手法は、eBay から取得した実データにおいて有効性が確かめられた [45]。

本研究において、我々は大規模なオンライン C2C (consumer-to-consumer, 消費者間商取引) 市場であるメルカリから取得したデータを解析する。これは日本最大のオンライン C2C 市場であり、2019 年の時点で、月間アクティブユーザー 1,300 万人、四半期取引総額 1,330 億円の規模である [53]。ここで、我々はユーザーの取引ネットワークに基づいて不正取引の解析を行なうが、これはユーザーの評価における不正を調べた先行研究とは対照的である [44–46, 48]。ネットワークに基づく不正検出アルゴリズムの先行研究の多くでは、ネットワークについての大域的な情報、たとえば連結成分や、コミュニティ、媒介中心性、k-core、確率伝播法によって決まる量などを用いていた [28–47, 49, 50]。その他の研究ではネットワークについての局所的な情報、たとえば次数や、三角形の数、ローカルクラスタリング係数などが用いられた [28–30, 34, 37, 44, 47, 48, 51, 52]。我々は局所的な情報、すなわち頂点と隣接頂点のつながり方、および隣接頂点同士のつながり方から計算できる特徴量を用いる。その理由は、局所的な特徴量の方が計算がより簡単かつ高速であり、サービスに実装することを考えても実用的であるためである。

2.2 材料と手法

2.2.1 データ

メルカリはオンライン C2C 市場のサービスであり、そこではユーザーは様々なアイテムを直接取引することができる。サービスは日本およびアメリカで運営されている。本研究では、日本の市場から得られたデータを用いた。通常取引に加えて、我々は以下の4種類の不正なアイテムを取引したユーザーに着目した：すなわち、fictive, underwear, medicine, weapon である。Fictive は存在しないアイテムを売りに出すことと定義される。Underwear は使用済み下着の取引を指す；これは倫理や衛生の観点からサービスによって禁止されている。Medicine は医薬品の取引を指す；これは法律によって禁止されている。Weapon は武器類の取引を指す；これは犯罪につながりうるためサービスによって禁止されている。サンプルした各種類のユーザーの数は、表 1 に示してある。

2.2.2 ネットワーク解析

我々はユーザーのつくる、有向・重み付きのネットワークを検討する。ここで、ユーザーはネットワークの頂点に、ユーザー間の取引は有向枝に対応する。枝の重みは、2ユーザー間でなされた取引の数をあらわす。今回我々は、サンプルした各数百の通常ユーザーおよび不正ユーザー（すなわち、少なくとも一回不正アイテムを売ったユーザー）をそれぞれ中心とする egocentric network を構築した。Egocentric network とは、中心となる頂点に隣接する頂点とその間の枝、および隣接頂点間の枝を含むネットワークである。

我々はそれぞれの中心頂点について、8つの指標を計算した。これらはすべて局所的な指標、すなわち、中心頂点と隣接する頂点間のつながり方までを情報として用いる指標である。

8つの指標のうち5つは、中心頂点のつながり方の情報のみを用いる。頂点 v_i の次数 (degree) k_i は、その頂点の隣人の数を示す。頂点 v_i の node strength [54] (重み付き次数 weighted degree) s_i は、その頂点の関わる取引の総数を示す。これらふたつの指標から、隣人一人あたりとの平均取引数 s_i/k_i を単独の指標として計算した。以上3つの指標では枝の向きは考慮しない。

頂点 v_i の sell probability SP_i は、枝の向きを考慮する指標で、 v_i が売り手として振る舞った隣人の割合を示す。詳細には、 SP_i は

$$SP_i = \frac{k_i^{\text{out}}}{k_i^{\text{in}} + k_i^{\text{out}}} \quad (1)$$

と定義される。ここで k_i^{in} は v_i の入次数、すなわち v_i が1回以上アイテムを買った隣接頂点の数、 k_i^{out} は v_i の出次数、すなわち v_i が1回以上アイテムを売った隣接頂点の数である。注意として、 v_i が隣接頂点 v_j に対する買い手でも売り手でもある場合、 v_j は v_i の入次数および出次数両方に1貢献する。それゆえ、 $k_i^{\text{in}} + k_i^{\text{out}}$ は次数 k_i と一般には等しくない。

重み付きの sell probability WSP_i は

$$\text{WSP}_i = \frac{s_i^{\text{out}}}{s_i^{\text{in}} + s_i^{\text{out}}} \quad (2)$$

と定義される。ここで、 s_i^{in} は v_i の重み付き入次数、すなわち買いの回数、 s_i^{out} は v_i の重み付き出次数、すなわち売りの回数である。

残り3つの指標は、中心頂点に関わる三角形に基づく。ローカルクラスタリング係数 C_i は、 v_i の周りの無向・重みなしの三角形の豊富さを定量化する [4]。これは v_i を含む無向・重みなしの三角形の数を $k_i(k_i - 1)/2$ で割った量として定義される。ローカルクラスタリング係数 C_i は0から1の間の値をとる。

我々は、ローカルクラスタリング係数の増加に貢献する三角形は、 v_i の特定の隣接頂点の周りに密集しているという仮説を立てた。そのような隣接頂点は v_i とともに、コミュニティともみなせるような重複した三角形のセットを形成しうる [55, 56]。よって我々の仮説は、中心頂点がコミュニティに関わる程度は、通常ユーザーと不正ユーザーで異なるということの意味する。この考えを定量化するため、triangle congregation m_i という量を導入する。これは、 v_i が関わる三角形のペアがもうひとつの頂点を共有する割合、すなわち

$$m_i = \frac{(v_i \text{ が関わる三角形のペアで、ほかにもうひとつの頂点を共有するものの数})}{\text{Tr}_i(\text{Tr}_i - 1)/2} \quad (3)$$

として定義される。ここで $\text{Tr}_i = C_i k_i(k_i - 1)/2$ は v_i が関わる三角形の数である。 m_i は0から1の間の値をとる。

モチーフ [57] と呼ばれる、3頂点の有向部分ネットワークの頻度は、通常ユーザーと不正ユーザーを区別しうる。特に、有向枝が構成する三角形を考えたとき、我々は feedforward 三角形 (図 1a) は自然だが、cyclic 三角形 (図 1b) はそうでないという仮説を立てた。なぜなら、feedforward 三角形の自然な解釈としては、出次数2の頂点は売り手を、出次数0の頂点は買い手を務める傾向にあるということになる。つまりユーザーの多くは、売り買い両方ではなく、売り手あるいは買い手のどちらかとして市場を使うということである。これに対して、cyclic 三角形が多数存在するということは、比較的多くのユーザーが売り買い両方に市場を使うということの意味する。我々は cycle probability CYP_i という量を用いた。

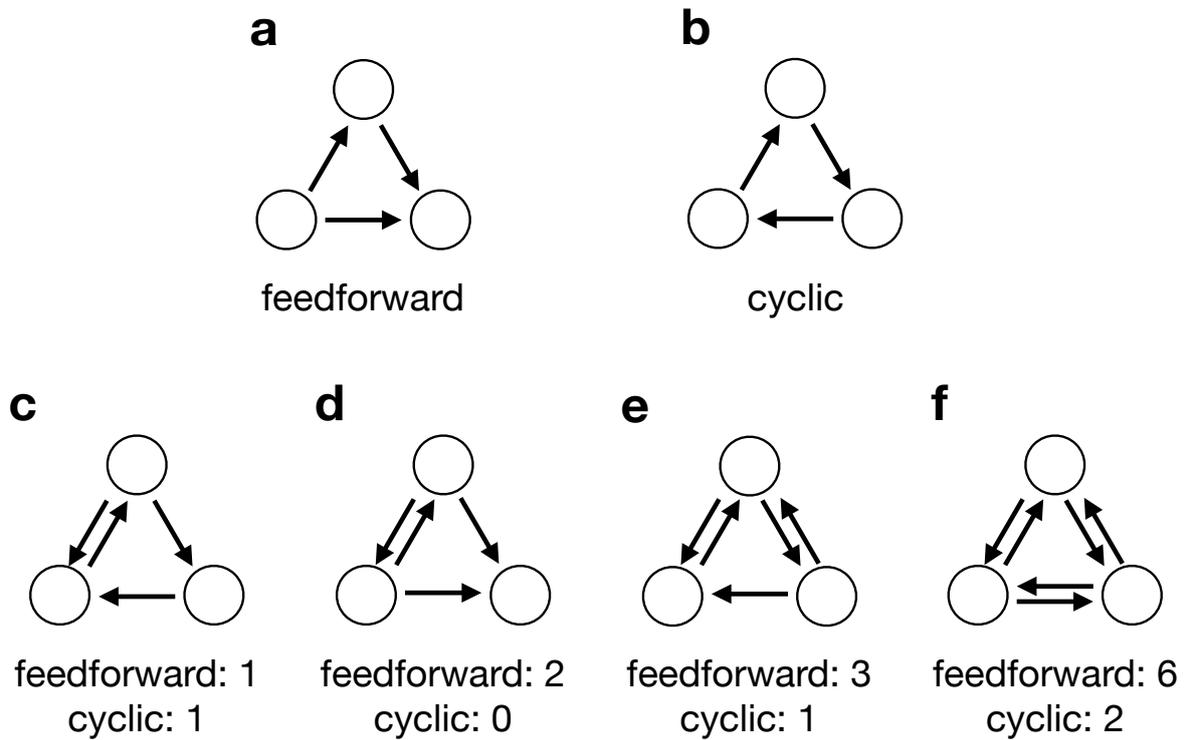


図 1. 有向三角形のパターンとその数え方。(a) Feedforward 三角形。(b) Cyclic 三角形。(c)–(f) 有向三角形と双方向の枝を含む 3 頂点パターン 4 種。図の下の数字はそのパターンが feedforward と cyclic 三角形の数にどのくらい貢献するかを示す。

これは

$$CYP_i = \frac{CY_i}{FF_i + CY_i} \quad (4)$$

と定義される。ここで、 FF_i と CY_i はそれぞれ v_i が関わる feedforward 三角形と cyclic 三角形の数である。 FF_i , CY_i , そして CYP_i の定義は、 v_i が関わる三角形が双方向の枝をもつ場合も有効である。たとえば図 1c の場合、3 つの頂点のいずれの関係も、1 つの feedforward 三角形と 1 つの cyclic 三角形を含んでいる。その他の双方向の枝をもつ場合に関しても図 1d-f に示してある。 CYP_i の計算では、枝の重みは考慮しなかった。

2.2.3 ランダムフォレストによる分類

局所的なネットワークの特徴に基づいて通常ユーザーと各種類の不正ユーザーとを分類するため、我々は scikit-learn [58] に実装されているランダムフォレスト [59–61] によって二値分類を行なった。これはアンサンブル学習を用いる手法、すなわち複数の分類器（ここでは決定木）を組み合わせることによって、過学習を避けつつ分類を行なう手法である。今回はひとつのランダムフォレスト分類器に、300 個の決定木を用いることとした。それぞれの決定木は、すべての学習サンプルから復元抽出によってランダムサンプリングされた学習サンプルに基づき構築される。また決定木の各節点で最適な分割を計算するには、すべての特徴量から候補となるものをランダムに抽出する。ひとつの決定木において、あるテストサンプルの陽性の確率を計算するには、テストサンプルが最終的に到達した葉を考える。その葉に到達した学習サンプルのうち陽性のサンプルの割合が、そのテストサンプルの陽性の確率となる。この陽性の確率を 1 から引いた値が、同じテストサンプルの陰性の確率となる。ランダムフォレストとしての陰性/陽性の確率は、ひとつひとつの決定木での陰性/陽性の確率を 300 個の木全体で平均することで得られる。陽性の確率が 0.5 より大きければ陽性、そうでなければ陰性と分類される。

今回は通常ユーザーの数が、どの種類の不正ユーザーの数よりも多い。そのため学習の際には、陰性（通常）サンプルの数と陽性（不正）サンプルの数とを均衡させるため、陰性サンプルのダウンサンプリングを行ない、通常と各不正種類とでサンプルの数を同数にした。その後、陰性/陽性あわせた各不正種類についてのサンプルセットを、75 % の学習サンプルセットと 25 % のテストサンプルセットとに分割した。このように構築された学習サンプルに基づき、300 個の決定木、すなわちひとつのランダムフォレスト分類器を作成した。そして、この分類器の分類性能をテストサンプルを用いて測定した。

分類器の性能を測るためのいくつかの指標を説明する。True positive rate は recall とも呼ばれ、陽性サンプル（不正ユーザー）のうち、分類器が正しく陽性と分類できたものの割合として定義される。False positive rate は、陰性サンプル（通常ユーザー）のうち、分類器が誤って陽性と分類したものの割合として定義される。Precision は、分類器が陽性と分類したサンプルのうち、真に陽性であるものの割合として定義される。以上 3 つの値はいずれも 0 から 1 の間の値をとる。

我々は分類器の性能を評価するため、receiver operating characteristic (ROC) 曲線と、precision-recall (PR) 曲線のふたつの指標を用いた。ROC 曲線を描くには、まずサンプルをそれらの陽性の確率によって降順で並べる。次に、順番に各サンプルを用いて、横軸に false positive rate を、縦軸に true positive rate をプロットする。そして、各点を区分的

に線形に結ぶことで、ROC 曲線が得られる。PR 曲線も同様に並べたサンプルで、横軸に recall を、縦軸に precision を描くことで得られる。正確な二値分類器では、ROC 曲線・PR 曲線ともに $(x, y) = (0, 1)$ 付近を通る。したがって、それぞれの曲線の area under curve (AUC, 曲線下面積) によって、分類器の性能を定量化することができる。AUC は 0 から 1 の間の値をとり、大きいほど分類器の性能が良いことを表す。

分類器における各特徴量の重要度を測るため、我々は permutation importance を用いた [62, 63]。この手法ではある特徴量の重要度は、テストセットのサンプル間でその特徴量をランダムに並べ替えたとき、学習済みの分類器の性能が低下する度合いによって測られる。この値が大きいほど、その特徴量が分類器の性能に貢献していることを表す。Permutation importance を計算するための評価指標として、我々は ROC 曲線の AUC を用いた。計算に際しては、各特徴量につき 10 回の異なる並べ替えを行ない、それらから計算した値の平均値を重要度として採用した。

ハイパーパラメータの最適化は、グリッドサーチとクロスバリデーション (10-fold) によって行なった。各決定木の最大深さ (scikit-learn における max_depth) は、3 から 10 の間の値を探索した。各節点での分割における特徴量の候補数 (max_features) は、3 から 6 の間の値を探索した。葉における最小サンプル数 (min_samples_leaf) は、1, 3, 5 を探索した。前述の通り、決定木の数 (n_estimators) は 300 とした。乱数発生器のシード (random_state) は 0 とした。その他のハイパーパラメータについては、scikit-learn version 0.22 のデフォルトの値を用いた。最適化の際の性能評価は、ROC 曲線の AUC を用いて行なった。

通常ユーザーのサンプリングのバイアスを避けるため、我々は 100 個のランダムサンプリングされたサンプルセットを用いてそれぞれランダムフォレスト分類器を作成し、個別に性能評価を行なった。

2.3 結果

2.3.1 記述統計

各ユーザーの種類について、次数の survival probability (ある値より大きな次数をもつ頂点の割合) を図 2a に示した。約 60 % の通常ユーザーが $k_i = 1$ であるのに対し、不正ユーザーはどの種類においても $k_i = 1$ のユーザーは 2 % 程度以下であった (表 1)。したがって、 $k_i = 1$ であるか $k_i \geq 2$ であるかは、通常ユーザーと不正ユーザーを区別するのに有用な情報であると期待される。次いで、 $k_i \geq 2$ のユーザーの次数分布も更に有用たりうることを考え、その survival probability を図 2b に示した。しかしながら、 $k_i = 1$ のユーザーの差ほどには顕著な差は見られなかった (表 1)。

表 1. 各種ユーザーの統計情報。第 1 列でたとえば $\text{mean}(A | B)$ という表記は、 B という条件のもとでの A の mean (平均値) を表す。また第 1 列で mean や median という表記がない場合は、表に示した数字はユーザーの数を意味する。

	Normal	Fictive	Underwear	Medicine	Weapon
Total	999	440	468	469	416
$k_i = 1$	587 (58.8%)	8 (1.8%)	3 (0.6%)	2 (0.4%)	5 (1.2%)
$\text{mean}(k_i k_i \geq 2)$	195.0	138.3	297.8	184.2	179.7
$\text{median}(k_i k_i \geq 2)$	77.5	61.0	170.0	97.0	86.0
$s_i = 1$	587 (58.8%)	8 (1.8%)	3 (0.6%)	2 (0.4%)	5 (1.2%)
$\text{mean}(s_i s_i \geq 2)$	365.1	153.3	325.3	198.1	199.4
$\text{median}(s_i s_i \geq 2)$	89.0	66.5	175.0	100.0	90.0
$s_i \geq 2$	412	432	465	467	411
$s_i/k_i = 1$	97 (23.5%)	97 (22.5%)	86 (18.5%)	156 (33.4%)	121 (29.4%)
$\text{mean}(s_i/k_i s_i/k_i > 1)$	1.413	1.135	1.055	1.066	1.092
$\text{median}(s_i/k_i s_i/k_i > 1)$	1.124	1.059	1.03	1.031	1.055
$k_i \geq 2$	412	432	465	467	411
$\text{SP}_i = 1$	157 (38.1%)	15 (3.5%)	21 (4.5%)	16 (3.4%)	17 (4.1%)
$k_i^{\text{out}} = 1$	118 (28.6%)	21 (4.9%)	2 (0.4%)	2 (0.4%)	9 (2.2%)
$s_i \geq 2$	412	432	465	467	411
$\text{WSP}_i = 1$	157 (38.1%)	15 (3.5%)	21 (4.5%)	16 (3.4%)	17 (4.1%)
$s_i^{\text{out}} = 1$	118 (28.6%)	14 (3.2%)	2 (0.4%)	2 (0.4%)	9 (2.2%)
$k_i \geq 2$	412	432	465	467	411
$C_i = 0$	118 (28.6%)	152 (35.2%)	108 (23.2%)	154 (33.0%)	128 (31.1%)
$\text{mean}(C_i C_i > 0)$	8.554×10^{-3}	8.348×10^{-3}	9.500×10^{-4}	2.231×10^{-3}	3.810×10^{-3}
$\text{median}(C_i C_i > 0)$	2.411×10^{-3}	2.039×10^{-3}	5.288×10^{-4}	6.494×10^{-4}	1.337×10^{-3}
$\text{Tr}_i \geq 2$	262	241	317	251	244
$m_i = 0$	17 (6.5%)	27 (11.2%)	54 (17.0%)	44 (17.5%)	32 (13.1%)
$m_i = 1$	12 (4.6%)	9 (3.7%)	4 (1.3%)	6 (2.4%)	11 (4.5%)
$\text{mean}(m_i m_i > 0)$	8.554×10^{-3}	8.348×10^{-3}	9.500×10^{-4}	2.231×10^{-3}	3.810×10^{-3}
$\text{median}(m_i m_i > 0)$	2.411×10^{-3}	2.039×10^{-3}	5.288×10^{-4}	6.494×10^{-4}	1.337×10^{-3}
$\text{FF}_i + \text{CY}_i \geq 1$	294	280	357	313	283
$\text{CYP}_i = 0$	234 (79.6%)	188 (67.1%)	222 (62.2%)	227 (72.5%)	202 (71.4%)
$\text{mean}(\text{CYP}_i \text{CYP}_i > 0)$	1.987×10^{-2}	7.367×10^{-2}	6.739×10^{-2}	8.551×10^{-2}	5.544×10^{-2}
$\text{median}(\text{CYP}_i \text{CYP}_i > 0)$	1.521×10^{-2}	4.481×10^{-2}	3.396×10^{-2}	3.822×10^{-2}	3.618×10^{-2}

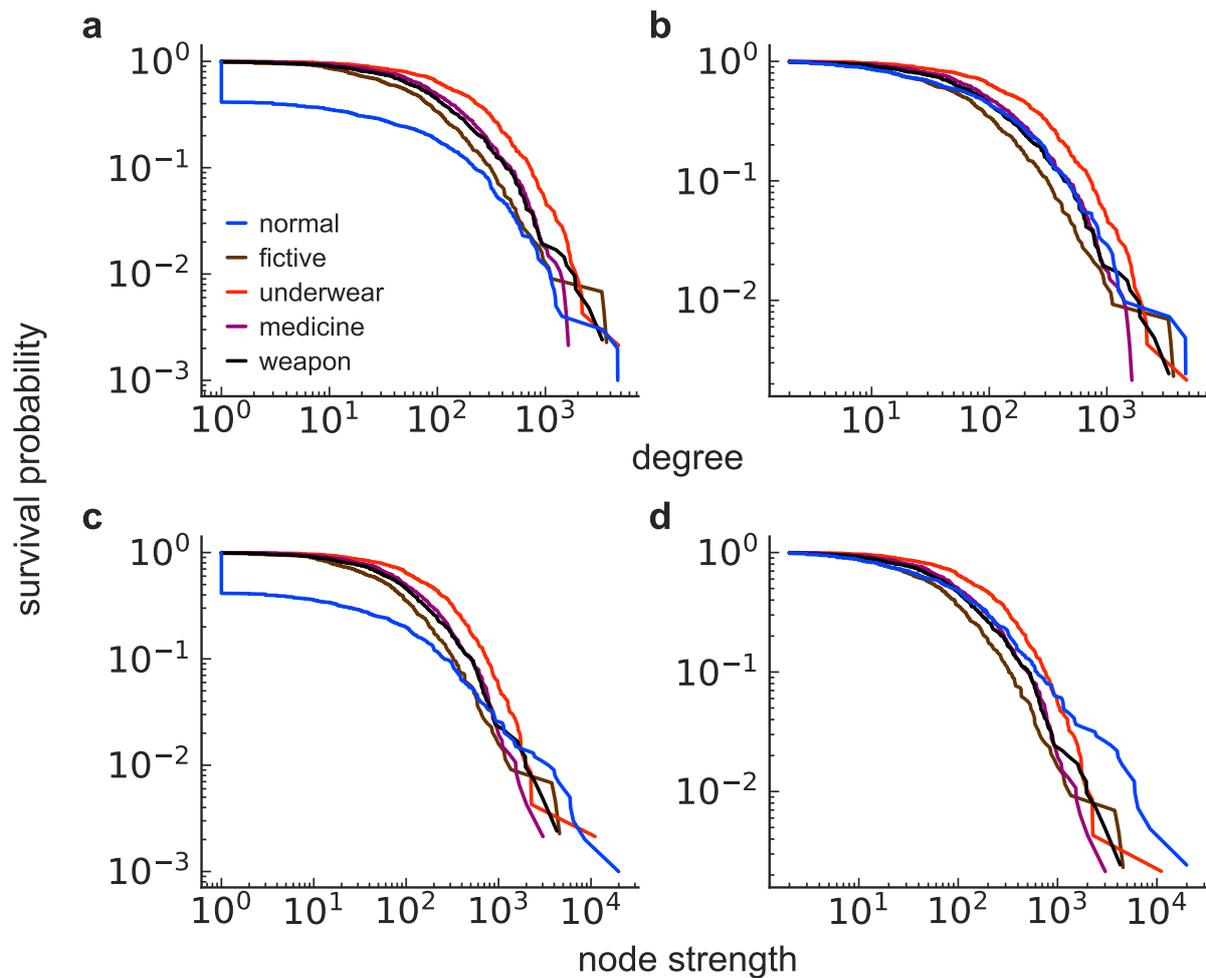


図 2. 各種ユーザーの度数。(a) 全頂点の度数 (k_i)。 (b) $k_i \geq 2$ の頂点の度数。(c) 全頂点の strength (s_i)。 (d) $s_i \geq 2$ の頂点の strength。

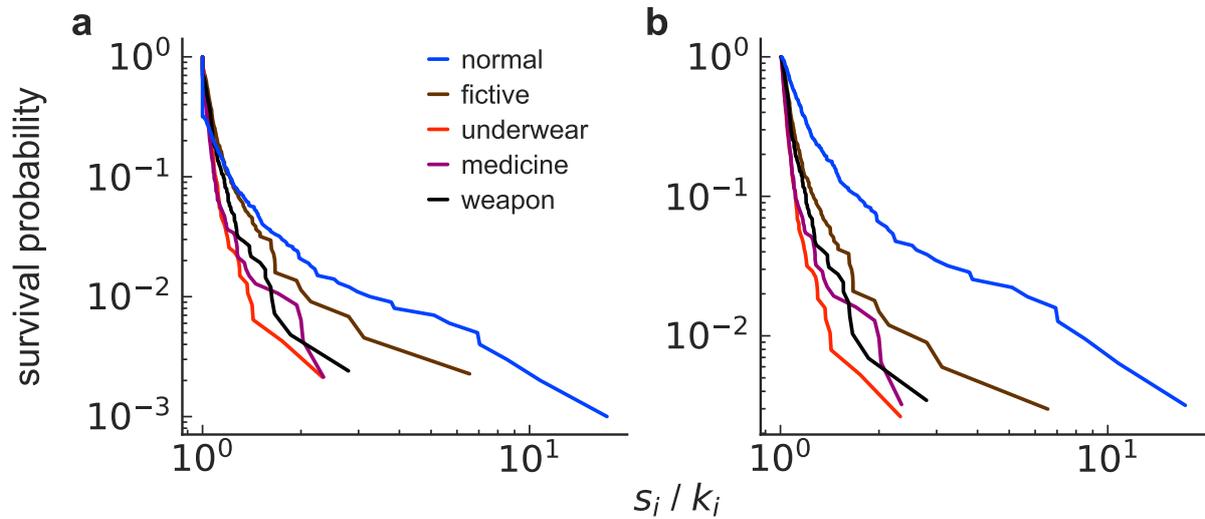


図 3. 各種ユーザーの隣人一人あたりの平均取引数。(a) 全頂点での s_i/k_i . (b) $s_i/k_i > 1$ の頂点の s_i/k_i .

各種ユーザーの node strength (重み付き次数) を図 2c に示した。重みなしの次数と同様に、多くの通常ユーザーが $s_i = 1$ であったが、不正ユーザーはそうではなかった。 $s_i \geq 2$ の条件のもとでの node strength の survival probability は、通常ユーザーと不正ユーザーとで顕著な差を示さなかった (図 2d、表 1)。

各種ユーザーの隣人一人あたりの平均取引数 (s_i/k_i) の分布を図 3a に示した。大部分の通常ユーザーでは $s_i/k_i = 1$ であった。これはかなりの数の通常ユーザーが隣人一人あたり 1 回だけ取引をしていることを意味する (表 1)。この結果は通常ユーザーの約 60% で $k_i = s_i = 1$ であるという結果と一致する。不正ユーザーではほとんどが $s_i/k_i > 1$ である。しかし全体的な分布としては、不正ユーザーが通常ユーザーよりもやや小さな値をもつ傾向が見える。この差は $s_i/k_i = 1$ のユーザーを除いた分布で顕著になる (図 3b、表 1)。したがって、特定の隣人との取引が少ないことが、不正ユーザーの特徴のひとつとして考えられる。

各種ユーザーの重みなし sell probability の分布を、図 4a に示した。通常ユーザーの分布は 0 と 1 付近にピークがあり、比較的多数の通常ユーザーが売り手または買い手のみに専念していることを示している。注意として、今回サンプルしたのは 1 回以上アイテムを売ったユーザーであるため、定義より sell probability は最低でも $1/(k_i^{\text{in}} + k_i^{\text{out}})$ である。したがって 0 付近のピークは、ユーザーが最低限の 1 回の売り以外には、全くあるいはほとんど売りを行なっていないことを意味する。対照的に、不正ユーザーの分布はどの種類でも比較的平坦である。重みなし sell probability と買い手・売り手の数の和 (sell probability の分母) の関係を、図 4b に示した。図中の点線上の頂点は、sell probability が $1/(k_i^{\text{in}} + k_i^{\text{out}})$ に等しく、その頂点が最小の出次数をもつ、すなわち $k_i^{\text{out}} = 1$ であることを意味する。つまり、この点線上のユーザーは隣人一人に対してを除き、すべて買い手として振る舞う。図 4b は、点線上のユーザーの大多数は通常ユーザーであることを示し、これは表 1 で定量的にも確かめられる。またほとんどの通常ユーザーは、sell probability = 1 の横線上 ($k_i \geq 2$ の通常ユーザーの 38.1 %) か点線上 (28.6 %) にのる。これは不正ユーザーではどの種類にも当てはまらない (表 1)。

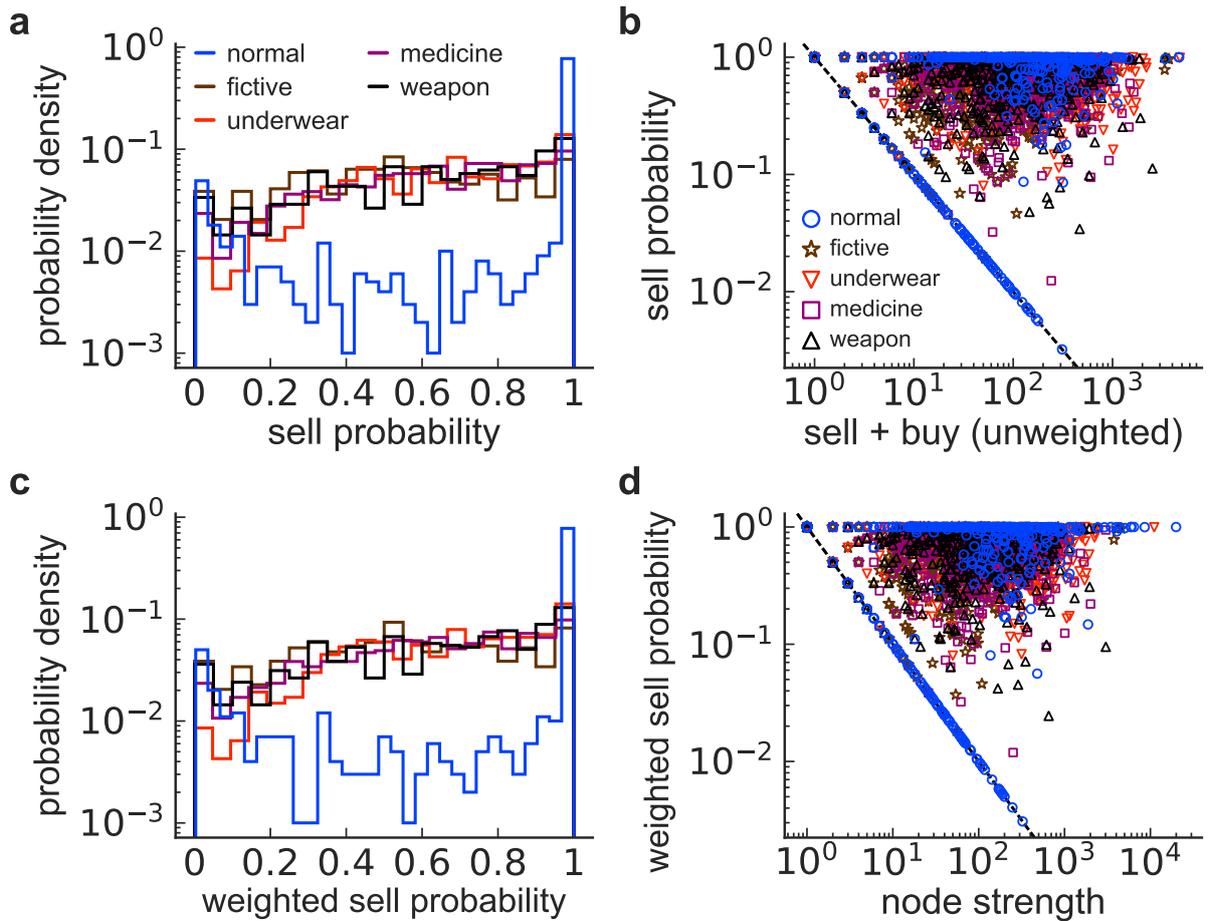


図 4. 各種ユーザーの sell probability. (a) 重みなし sell probability の分布。(b) 買い手・売り手の数の和と重みなし sell probability の関係。(c) 重み付き sell probability の分布。(d) 買い取引・売り取引の数の和と重み付き sell probability の関係。(b) と (d) における点線は、それぞれ $1/(k_i^{\text{in}} + k_i^{\text{out}})$ と $1/(s_i^{\text{in}} + s_i^{\text{out}})$ を示す。

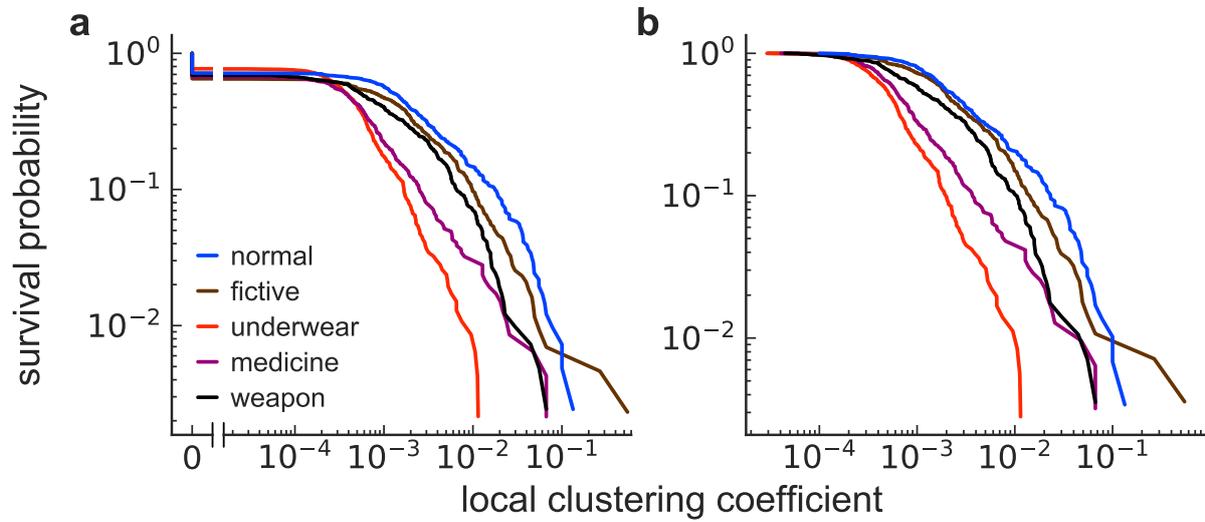


図 5. 各種ユーザーのローカルクラスタリング係数。(a) Survival probability. (b) $C_i > 0$ に限定した survival probability.

ローカルクラスタリング係数の survival probability を図 5a に示した。注意として、 $k_i = 1$ の場合 C_i は定義できないため、この解析は $k_i \geq 2$ のユーザーに限定した。 $C_i = 0$ のユーザーの数は、通常ユーザーと不正ユーザーとで顕著な違いはなかった (表 1)。 $C_i > 0$ に限定した survival probability を図 5b に示した。通常ユーザーは不正ユーザーより大きな C_i の値をもつ傾向が見えるが、それほど強くはない (表 1)。

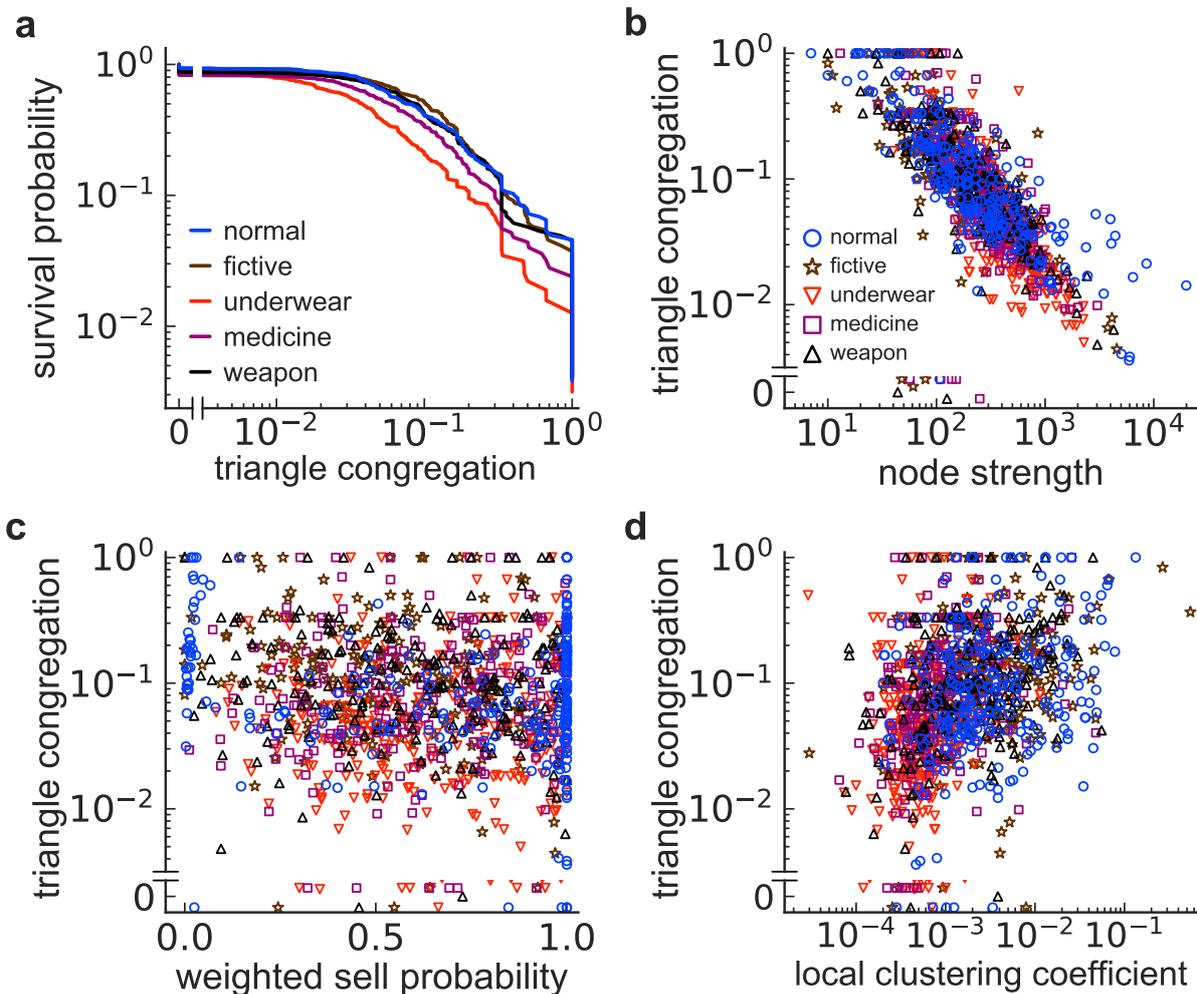


図 6. 各種ユーザーの triangle congregation. (a) Survival probability. (b) Triangle congregation m_i と node strength の関係。 (c) m_i と重み付き sell probability の関係。 (d) m_i とローカルクラスタリング係数の関係。

Triangle congregation の survival probability を図 6a に示した。我々の仮説に反して、通常ユーザーと不正ユーザーの分布に顕著な違いはなかった。Node strength が小さいとき (図 6b) と、ローカルクラスタリング係数が大きいとき (図 6d) に、triangle congregation は大きい傾向がある。重み付き sell probability との関係はまちまちである (図 6c)。しかしながら、通常ユーザーと不正ユーザーとではっきりした違いは見られなかった (表 1 参照)。

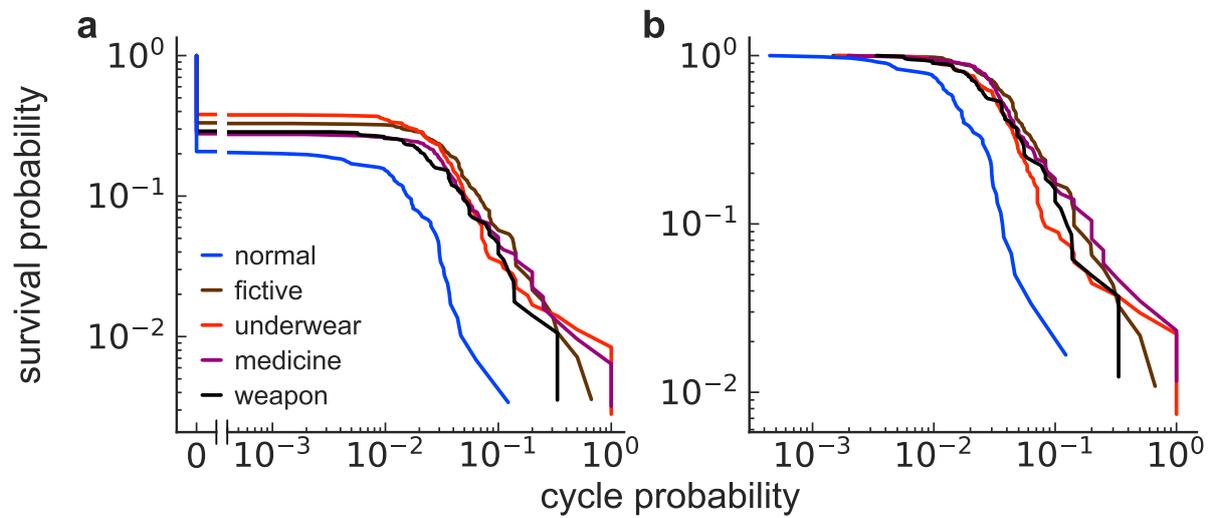


図 7. 各種ユーザーの cycle probability. (a) Survival probability. (b) $CYP_i > 0$ に限定した survival probability.

Cycle probability の survival probability を図 7a に示した。どの種類のユーザーでも、大多数は $CYP_i = 0$ であった (表 1)。 $CYP_i = 0$ のユーザーを除いた場合、通常ユーザーほどの種類の不正ユーザーよりも小さな CYP_i の値を持つ傾向が見られた (図 7b、表 1)。

2.3.2 ユーザーの分類

前項で解析した8個の指標に基づき、我々は分類器に使用する12個の特徴量を定義した。分類器の目的は通常ユーザーと不正ユーザーとを識別することである。第1の特徴量は、次数 $k_i = 1$ か $k_i \geq 2$ かの二値である。第2の特徴量は、node strength $s_i = 1$ か $s_i \geq 2$ かの二値である。第3の特徴量は、 s_i/k_i で、これは1以上の値をとる実数である。第4の特徴量は、重みなし sell probability $SP_i = 1$ か $SP_i < 1$ かの二値である。第5の特徴量は、 $SP_i = 1/(k_i^{\text{in}} + k_i^{\text{out}})$ か $SP_i > 1/(k_i^{\text{in}} + k_i^{\text{out}})$ か、すなわち $k_i^{\text{out}} = 1$ か $k_i^{\text{out}} > 1$ かの二値である。第6の特徴量は、 SP_i で、これは0から1の間の値をとる実数である。第7の特徴量は、重み付き sell probability $WSP_i = 1$ か $WSP_i < 1$ かの二値である。第8の特徴量は、 $WSP_i = 1/(s_i^{\text{in}} + s_i^{\text{out}})$ か $WSP_i > 1/(s_i^{\text{in}} + s_i^{\text{out}})$ か、すなわち $s_i^{\text{out}} = 1$ か $s_i^{\text{out}} > 1$ かの二値である。第9の特徴量は、 WSP_i で、これは0から1の間の値をとる実数である。第10の特徴量は、ローカルクラスタリング係数 C_i で、これは0から1の間の値をとる実数である。次数 $k_i = 1$ のときは、ローカルクラスタリング係数は定義できない。この場合、 $C_i = -1$ とした。第11の特徴量は、triangle congregation m_i で、これは0から1の間の値をとる実数である。頂点 v_i が1つ以下の三角形にしか関わっていないときは、 m_i は計算できない。この場合、 $m_i = -1$ とした。第12の特徴量は、cycle probability CYP_i で、これは0から1の間の値をとる実数である。頂点 v_i が feedforward 三角形にも cyclic 三角形にも関わっていないときは、 CYP_i は定義できない。この場合、 $CYP_i = -1$ とした。

これら12の特徴量をすべて使用して、fictive 種類の不正ユーザーを分類した結果のROC曲線とPR曲線を、図 8a, b にそれぞれ示した。細線はそれぞれが100個の分類器のうちの一つに対応する。太線は100本の細線の平均である。点線はランダムな分類に相当する。その他の不正種類の分類結果についても、図 8c-h に示した。図 8からは、分類性能は良好であることが伺える。定量的には、AUC はどの不正種類に対しても0.94を上回った(表 2)。

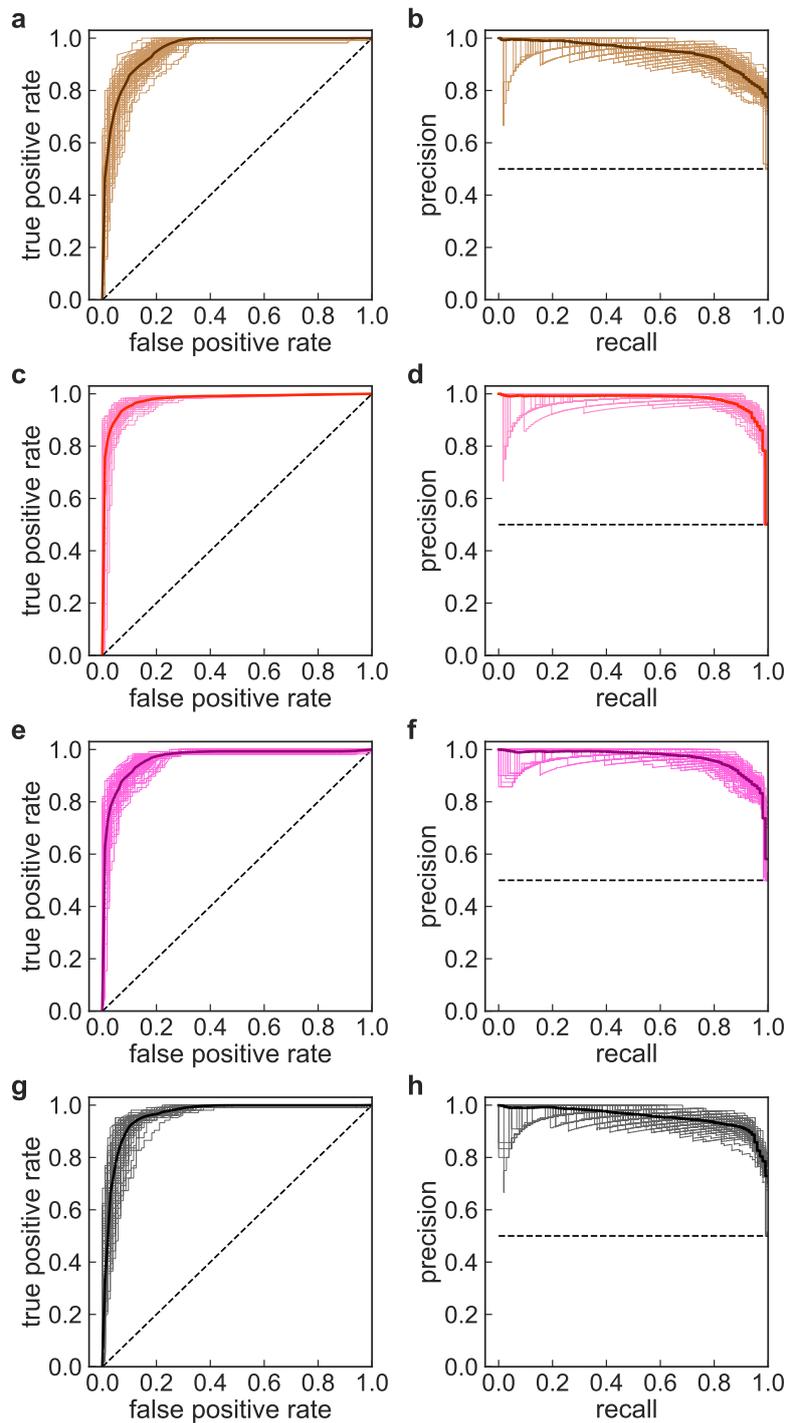


図 8. 通常ユーザーと不正ユーザーとを分類したときの ROC および PR 曲線。細線はそれぞれが 100 個の分類器のうちの一つに対応する。太線は 100 本の細線の平均である。点線はランダムな分類に相当する。
 (a) Fictive ユーザーを分類したときの ROC 曲線。(b) Fictive ユーザーを分類したときの PR 曲線。
 (c) Underwear ユーザーを分類したときの ROC 曲線。(d) Underwear ユーザーを分類したときの PR 曲線。
 (e) Medicine ユーザーを分類したときの ROC 曲線。(f) Medicine ユーザーを分類したときの PR 曲線。
 (g) Weapon ユーザーを分類したときの ROC 曲線。(h) Weapon ユーザーを分類したときの PR 曲線。

表 2. 分類器の AUC. 作成した 100 個の分類器から、平均と標準偏差を計算した。

		Fictive	Underwear	Medicine	Weapon
12 features	ROC	0.955 ± 0.011	0.978 ± 0.005	0.967 ± 0.009	0.963 ± 0.010
	PR	0.949 ± 0.015	0.978 ± 0.010	0.967 ± 0.011	0.956 ± 0.015
9 features	ROC	0.941 ± 0.015	0.964 ± 0.011	0.957 ± 0.011	0.956 ± 0.013
	PR	0.928 ± 0.024	0.950 ± 0.021	0.948 ± 0.021	0.937 ± 0.026

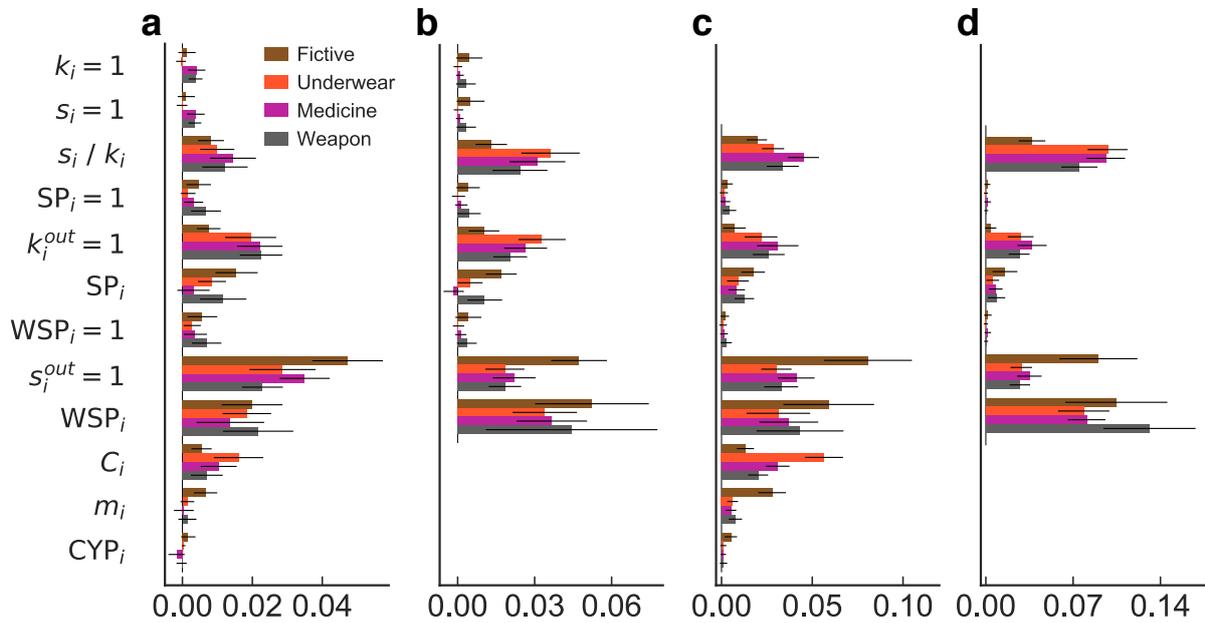


図 9. 分類器に用いた特徴量の permutation importance. (a) 12 個の特徴量。(b) 9 個の特徴量。(c) 10 個の特徴量。(d) 7 個の特徴量。太棒は 100 個の分類器の平均を表す。エラーバーは標準偏差である。

不正種類ごとの各特徴量の重要度を、図 9a に示した。重要度の傾向は、各不正種類でおおよそ似たものとなった。図 9a は、隣人一人あたりの平均取引数 (s_i/k_i)、売り手が最小限かどうか ($k_i^{out}=1$)、売り取引が最小限かどうか ($s_i^{out} = 1$)、重み付き sell probability (WSP_i) が、最も重要な特徴量群であることを示している。前項で見た記述統計とあわせると、小さい s_i/k_i 、 $k_i^{out} \neq 1$ 、 $s_i^{out} \neq 1$ 、中程度の WSP_i が、そのユーザーが不正ユーザーであることを示す特徴であると考えられる。

図 9a はまた、三角形に基づく特徴量 (C_i , m_i , CYP_i) は、分類性能に強くは貢献していないことを示唆している。これらの特徴量のみが隣人間のつながり方を必要とするため、これらを除いて同程度の分類性能を達成できれば、実用的には有益である；つまり、必要となる情報は中心頂点のつながり方のみになる。この可能性を探るため、我々は 12 個の特徴量のうち、隣人間のつながり方を用いない 9 個の特徴量のみを用いて分類器を構築した。不正ユーザーを分類したときの ROC 曲線と PR 曲線の AUC の平均値を、表 2 に示した。またこのときの ROC 曲線と PR 曲線を、図 10 に示した。すべての特徴量を用いたときよりやや性能は下がるものの、AUC はすべて 0.92 を超え、良好な性能を保った。このとき用いた 9 個の特徴量の permutation importance を、図 9b に示した。傾向は 12 個の特徴量の時とおおよそ同様であったが、 WSP_i の重要度がより増した。

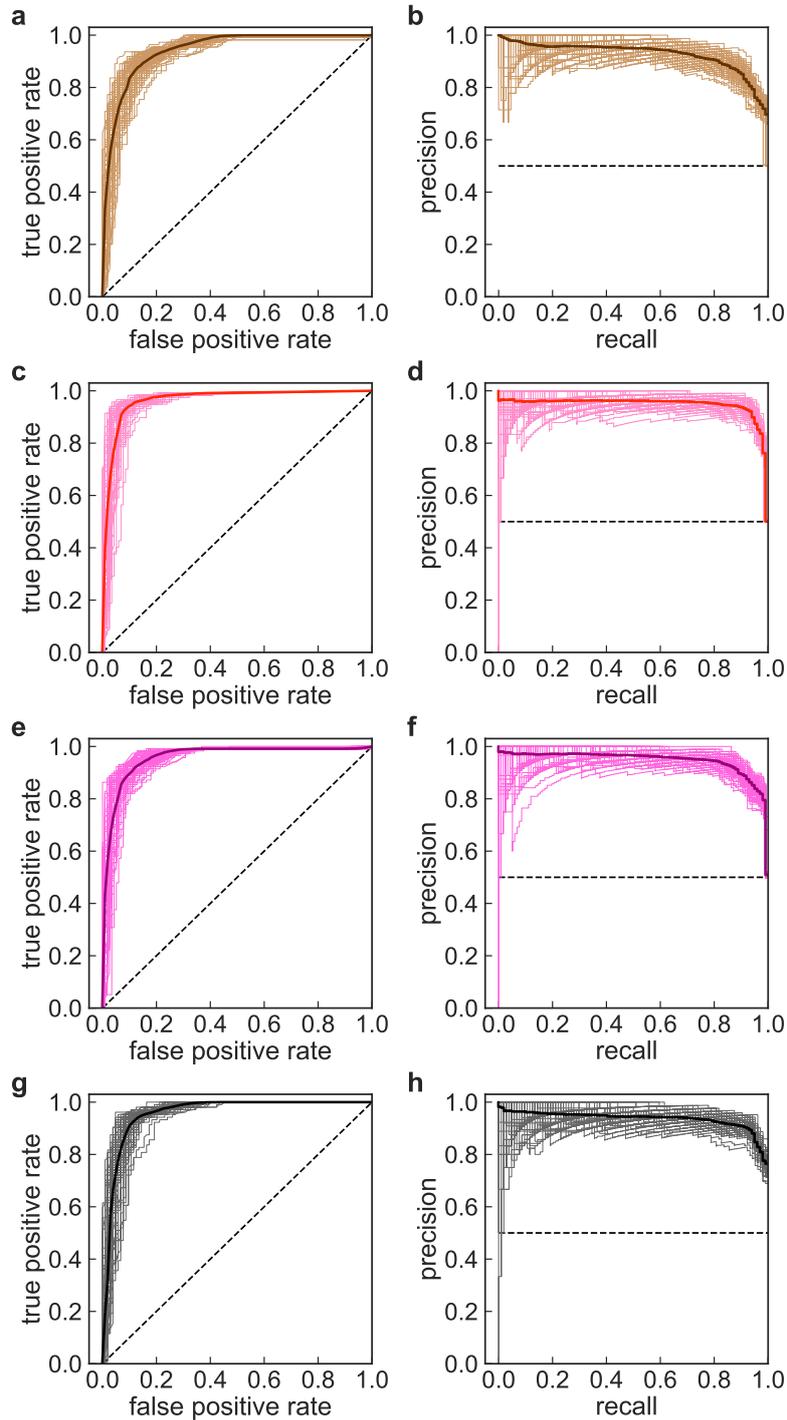


図 10. 隣人間のつながり方を使わずに分類したときの ROC および PR 曲線。(a) Fictive ユーザーを分類したときの ROC 曲線。(b) Fictive ユーザーを分類したときの PR 曲線。(c) Underwear ユーザーを分類したときの ROC 曲線。(d) Underwear ユーザーを分類したときの PR 曲線。(e) Medicine ユーザーを分類したときの ROC 曲線。(f) Medicine ユーザーを分類したときの PR 曲線。(g) Weapon ユーザーを分類したときの ROC 曲線。(h) Weapon ユーザーを分類したときの PR 曲線。細線はそれぞれが 100 個の分類器のうちの一つに対応する。太線は 100 本の細線の平均である。点線はランダムな分類に相当する。

表 3. $k_i = 1$ のユーザーを除いた分類器の AUC. 100 個の分類器から、平均と標準偏差を計算した。

		Fictive	Underwear	Medicine	Weapon
10 features	ROC	0.925 ± 0.016	0.950 ± 0.013	0.954 ± 0.012	0.916 ± 0.019
	PR	0.923 ± 0.019	0.950 ± 0.018	0.954 ± 0.016	0.911 ± 0.023
7 features	ROC	0.886 ± 0.020	0.921 ± 0.015	0.933 ± 0.014	0.899 ± 0.020
	PR	0.874 ± 0.027	0.901 ± 0.021	0.928 ± 0.019	0.880 ± 0.028

通常ユーザーの半分以上が $k_i = 1$ であり、これは不正ユーザーではどの種類でもごくわずかであった (表 1)。これにより通常/不正ユーザーの分類が簡単になり、性能の向上 (AUC の増加) につながっているかもしれない。これを調査するため、 $k_i = 1$ のユーザーを除いたサブデータを用いて分類を行なった (表 1)。この除外により陰性サンプルと陽性サンプルとがほぼ同数になったため、アンダーサンプリングは行わなかった。代わりに、100 個の異なる学習/テストセットへの分割をもとに、分類器を構築した。特徴量が ($k_i = 1$ かどうかと、 $s_i = 1$ かどうかを除いた) 10 個または 7 個のときの AUC を、表 3 に示した。また ROC および PR 曲線を、それぞれ図 11 および図 12 に示した。除外を行わなかった場合と比較して、十分競合できる程度の AUC を保った (表 2)。このときの各特徴量の重要度を、図 9c, d に示した。除外を行わなかった場合と比較して、重要とされる特徴量に顕著な違いは見られなかった。

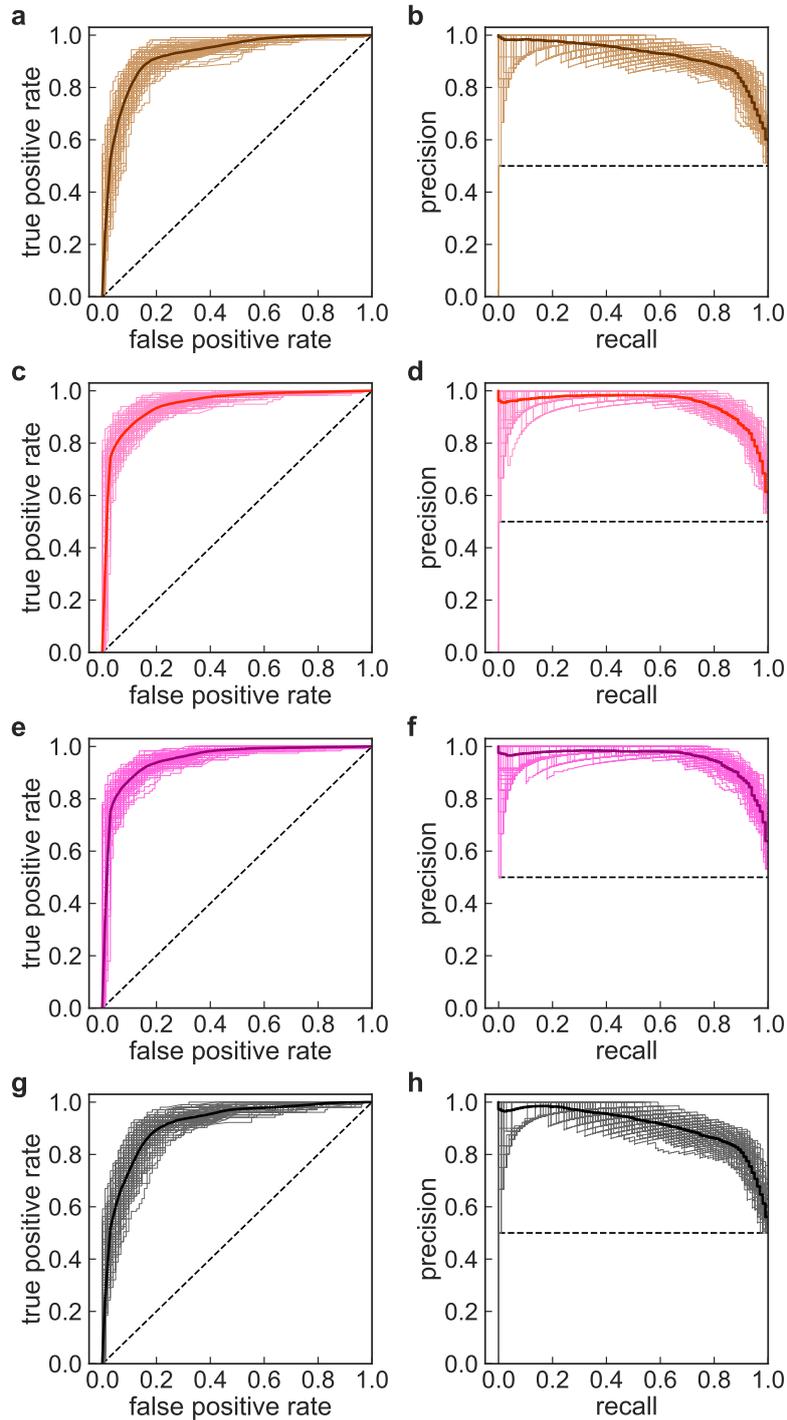


図 11. $k_i = 1$ のユーザーを使わずに分類したときの ROC および PR 曲線。(a) Fictive ユーザーを分類したときの ROC 曲線。(b) Fictive ユーザーを分類したときの PR 曲線。(c) Underwear ユーザーを分類したときの ROC 曲線。(d) Underwear ユーザーを分類したときの PR 曲線。(e) Medicine ユーザーを分類したときの ROC 曲線。(f) Medicine ユーザーを分類したときの PR 曲線。(g) Weapon ユーザーを分類したときの ROC 曲線。(h) Weapon ユーザーを分類したときの PR 曲線。細線はそれぞれが 100 個の分類器のうちの一つに対応する。太線は 100 本の細線の平均である。点線はランダムな分類に相当する。

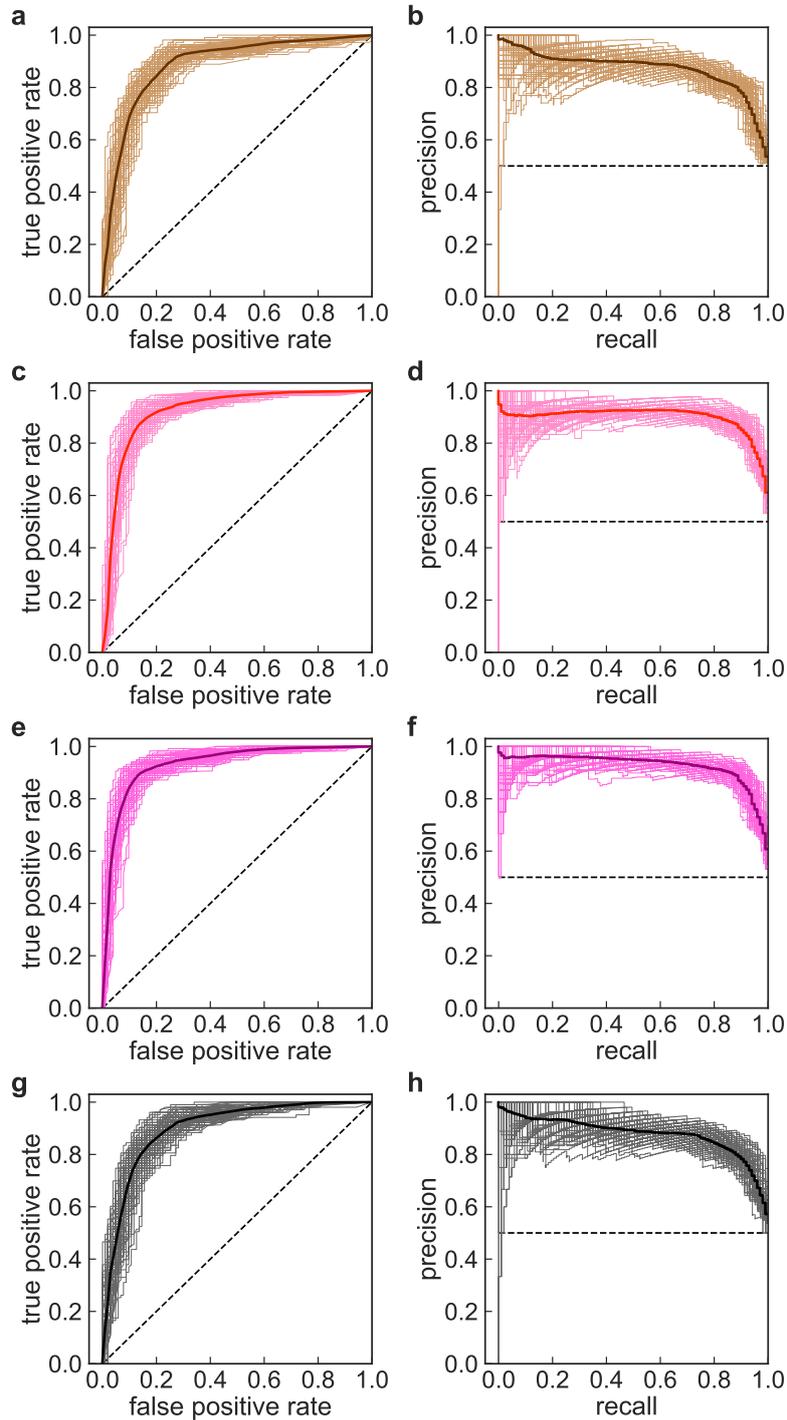


図 12. $k_i = 1$ のユーザーと隣人間のつながり方を使わずに分類したときの ROC および PR 曲線。
 (a) Fictive ユーザーを分類したときの ROC 曲線。(b) Fictive ユーザーを分類したときの PR 曲線。
 (c) Underwear ユーザーを分類したときの ROC 曲線。(d) Underwear ユーザーを分類したときの PR 曲線。
 (e) Medicine ユーザーを分類したときの ROC 曲線。(f) Medicine ユーザーを分類したときの PR 曲線。
 (g) Weapon ユーザーを分類したときの ROC 曲線。(h) Weapon ユーザーを分類したときの PR 曲線。細線
 はそれぞれが 100 個の分類器のうちの一つに対応する。太線は 100 本の細線の平均である。点線はランダ
 ムな分類に相当する。

2.4 考察

以上のように我々は、ユーザーがつくるネットワークの特徴による分類器を用いて、通常ユーザーと不正ユーザーとを高い確度 ($AUC \approx 0.95-0.98$) で判別できることを示した。特徴量を考えるにあたっては、中心となるユーザーの周りの、局所的な取引ネットワークの情報のみを用いた。このようにした理由は、ユーザーの数が非常に多いため、取引ネットワーク全体の情報を用いる必要がないのであれば、その方が実用的であるためである。また、中心ユーザーの隣人間のつながり方の情報を用いずに、中心ユーザーのつながり方の情報のみを用いた場合にも、 $AUC \approx 0.93-0.96$ という性能を達成できた。これにより、より少ない量のデータで分類を行なうことができるため、オンラインに実装する際には有利である。

本研究では分類器のアルゴリズムとしてランダムフォレストを使用した。これは恣意的な選択である。別のアルゴリズムを使用して、さらなる分類性能の向上を図ることもできるが、これは今後の課題である。その他の課題として、今回の結果が汎化できるかを見ることが挙げられる。つまり、転売チケットや、アダルト商品、盗品など他の不正アイテム、また他のプラットフォームなどに適用したときに、同様の結果を得ることができるかということである。特に、ある種類の不正ユーザーのサンプルで学習を行なった分類器によって、他の種類の不正ユーザーも検出することができれば、未知の種類の不正取引を検出するのに有効な手段となりうる。関連する課題として、異なる種類の不正取引を、単一の陽性としてまとめて学習した場合の性能を調べることが挙げられる。

ネットワークを用いた不正検出の先行研究では、ネットワークの大域的な特性を用いたものも局所的な特性を用いたものもあった。ネットワークの大域的な特性とは、個々の頂点に関する量を計算するのにネットワーク全体の構造を必要とするものであり、例として、連結成分 [28,31,43]、媒介中心性 [28-30]、確率伝播法によって決まる量 [34,35,41,44,45,47,49,50]、コミュニティなどの密な部分グラフ [28,32,33,36-39]、k-core [40,46] などがある。これらの手法は高い分類性能を発揮するものの、ネットワーク全体に関する情報を必要とする。ネットワークが大規模であったり急速に拡大するものであったりする場合、そのような全体のデータは得にくく、そのため計算速度やメモリ要求量、頂点や枝の情報の正確さなどの点で不利になりうる。他の手法は代わりに、ネットワークの局所的な特性、たとえば有向や重み付きの場合を含めた次数 [28-30,34,37,44,47,48,51,52] や、三角形や四角形の量 [34,51] などを用いた。ネットワークの局所的な量を使用することは、中心となる頂点が与えられれば迅速に計算が行えるため、産業において、特にサンプルされたユーザーに対して有利になることが見込まれる。局所的な特性に着目したもう一つの理由として、計算上の都合により

ネットワーク全体の構造を入手できなかったことがある。局所的な特性に加えて大域的な特性を使用することは分類性能を向上させうるが [37]、しかし局所的な特性のみを用いた本手法でも、大域的な特性を用いた手法の性能 (ROCAUC=0.880–0.986 [28, 31, 34, 35, 49, 50]) と同程度の性能を得ることができた。

本研究と同じメルカリのデータを用いた先行研究では、専門業者ではない望ましい常習ユーザー (売り手) と、望ましくない業者ユーザーとを分類することを目的としていた [64]。この論文 [64] の著者らは、ユーザーのプロフィール、アイテムの説明、一日の購入数などその他の行動についての情報を用いた。対照的に、我々はユーザーのつくるネットワークの局所的な特徴に着目した (ただし、[64] でも WSP_i に似た量は使われた)。加えて、[64] の著者らが不正取引を単一の広い区分と見ているのに対して、我々は特定の種類の不正取引を対象とした。ネットワークに関する特徴とネットワーク以外の特徴を組み合わせることで、さらなる性能の向上が実現されうるが、これも今後の課題である。

3 引用関係ネットワークの解析

本章の要旨

学術論文などのデータが蓄積されてきたことによって、文献の引用関係や共著関係などから科学そのものについての社会性の研究を行なう、science of science と呼ばれる分野が発展してきている。学術論文では、被引用数の多い文献の典型的な履歴として、出版直後に着目され被引用数が増加していき、その後新規性の低下などにより漸減していく、というものが考えられる。しかし、出版直後にはほとんど引用されず、あるときから急激に引用されるようになる、delayed recognition (DR) 文献と呼ばれるものも存在する。近年の検出手法によると、このような DR 文献はごく少数の例外ではなく、多数存在することが示唆されている。だが、既存研究は個別の事例の解析にとどまり、多数の DR 文献の間関係性は未だ明らかではない。本研究で我々は、DR 文献の間関係性をネットワークの観点から解析した。我々は物理分野のデータセットから抽出した DR 文献 300 本を用いて、それら間の引用関係ネットワーク、および共引用関係ネットワークを構築した。また、それらと同様の出版年と被引用数をもつランダム文献セットを作成、同様にネットワークを構築し、指標の比較を行なった。DR 文献セットはランダム文献セットと比較して、引用関係ネットワークではつながりが弱い、共引用関係ネットワークではつながりが強い傾向が見られた。また DR 文献セットから構築した共引用関係ネットワークをコミュニティに分割したところ、各コミュニティに属する文献で被引用履歴の同期が見られた。

3.1 導入

科学に携わる研究者らの活動によって、科学的成果、およびそれをまとめた学術論文などの文献が日々生産されてきた。これらの文献が蓄積され、また電子化などにより利用しやすくなったことから、大量の文献を俯瞰することでさらなる知識を生み出すことが可能になってきた [65]。たとえば文献の引用関係や共著関係などから、科学そのものにおける社会性の研究を行なう science of science と呼ばれる分野が興ってきており、研究の評価や方策決定への応用が期待されている [3]。

文献の引用は、学術論文などにおいて先行する文献との関連を記述する標準的な方法である。学術論文では各文献の被引用数を分布として見ると、比較的少数の文献が多数の引用を獲得していることが知られている [66]。被引用数が多い文献のたどる被引用の履歴としては、出版直後から着目・引用され、そのため人目につきやすくなり、更に引用数が増加する、という過程が考えられる [6, 7, 67]。実際に、出版直後に被引用数のピークを迎え漸減するもの、定期的に引用され続けるもの、徐々に被引用数が増えていくものなどが大多数であることが、実データの解析によって示唆されている [68]。

しかし、被引用数が多い文献の中にも、出版直後にはほとんど引用されず、ある時から急激に引用されるようになったものが存在する。このような文献および現象は、引用関係のダイナミクスや科学の発展を理解する上で重要であると考えられ、delayed recognition や Sleeping Beauty 文献などと呼ばれ着目されてきた [69–73]（本稿では delayed recognition (DR) 文献と呼ぶ [74]）。近年の検出手法の開発により、DR 文献はそれまで考えられてきたようなごく少数の例外ではなく、多数存在することが示唆されている [75]。しかし、その発生機序は未だ明らかではない。

文献の関係性を見る別の方法として、共引用がある [76]。これは、あるひとつの文献内で引用されている文献同士の間には、共引用の関係があるとして定義されるものである。複数の文献があるふたつの文献を引用しているという状況もありうるため、共引用関係には重みを定義することができる。したがって、あるふたつの文献間の共引用関係（の重み）は経時的に、すなわち新たな文献が出版されるたびに变化しうる。これは引用関係とは対照的である；あるふたつの文献間の引用関係は、引用する側の文献が出版された時点から变化することはない。

DR 文献に関する研究においても、共引用関係の解析は用いられてきた。それら先行研究の主な目的は、ある DR 文献が着目されるきっかけとなったひとつの文献 [77–81]、あるいは複数の文献 [82–85] を探すことであった。すなわちこれらの研究で解析されたのは、DR

文献と、それと同時に引用されているが DR 文献とはみなされない文献との関係性であった。またこれらの研究は、個別の DR 文献についてそれぞれ事情を検討するものであった。したがって、多数の DR 文献が検出できるようになったにもかかわらず、それら DR 文献一般の性質については、定量的な研究は未だ乏しい。

本研究では、DR 文献の間関係性をネットワークの観点から解析する。DR 文献のセットから引用関係ネットワーク、および共引用関係ネットワークを作成し、その性質を各種指標から考察する。また、ランダムな文献のセットから構築したネットワークでの結果と比較することで、DR 文献に特有の性質を定量的に示すことを目的とする。

3.2 材料と手法

3.2.1 データ

解析対象として、American Physical Society (APS) から取得した、Physical Review 系列の雑誌に掲載された物理学分野の文献データを用いた*2。このデータは 1893 年から 2016 年までに出版された 596,786 本の文献を含み、それぞれが ID、タイトル、掲載雑誌、出版日、著者、引用文献などの情報をもつ。注意として、引用関係はデータセット内部で完結している。すなわち、このデータ中に含まれない文献を本来引用していても、その情報は考慮されていない。データは 2018 年 1 月に取得した。

3.2.2 文献サブセットの構築

我々は上記の文献データセットから、解析対象とする DR 文献のサブセットを抽出した。まずデータに含まれる文献の DR 文献らしさを定量化するため、 B スコアという量を用いた [75]。これは各文献に対して年ごとの被引用数から計算される量で、

$$B = \sum_{t=0}^{t_m} \frac{l_t - c_t}{\max\{1, c_t\}} \quad (5)$$

と定義される。ここで、 $l_t = \frac{c_{t_m} - c_0}{t_m} \cdot t + c_0$ 、 t はその文献が出版されてからの経過年数（出版年で $t = 0$ ）、 t_m はその文献の年ごとの被引用数がピークに達したときの経過年数、 c_t は年 t におけるその文献の被引用数である。被引用数のピークを迎えるまでが長いほど、そしてそのピークが鋭いほど、スコアは大きくなる。したがってこのスコアが大きいほど、その文献は DR 文献らしいと考えられる。このスコアで上位 300 本の文献を、今回解析対象とする DR 文献セットとした。

また DR 文献セット特有の性質を検討するため、比較対象となるランダム文献セットを構築した。この構築は、上で選んだ 300 本の DR 文献それぞれと同程度の出版年と被引用数をもつ文献 300 本を、全データからランダムに抽出することで行なった。その疑似コードを以下の 1 に示した。

*2 <https://journals.aps.org/datasets>

Algorithm 1 ランダム文献セットの構築

```
input: set_DR_papers           ▷ DR 文献 300 本のセット
set_random_papers = (empty set)   ▷ 出力セット
least_choices = 3                 ▷ ランダム抽出の選択肢数の下限
year_range = 10                   ▷ 出版年の許容範囲
cited_range = 5                   ▷ 被引用数のランクの許容範囲 (可変)

for DR_paper in set_DR_papers do
  year_DR = (the year when DR_paper was published)
  cited_DR = (the number of citations of DR_paper)
  flag = 0
  ▷ 選択肢: 出版年と被引用数のランクが許容範囲内で、DR 文献セットに含まれない未抽出の文献
  while flag == 0 do:
    set_choices = (set of papers that
                   was published in year_DR ± year_range
                   and was cited cited_DR ± cited_range times
                   and not in set_DR_papers
                   and not in set_random_papers
                   )

    if length(set_choices) ≥ least_choices then
      set_random_papers += (random choice from set_choices)   ▷ ランダムにひとつ出力に追加
      flag = 1
    else
      ▷ 選択肢が少なければ、被引用数のランクのレンジを大きくしてやり直し
      citation_range += 5

return set_random_papers
```

この手続きを 100 回行ない、100 個の独立したランダム文献セットを構築した。

3.2.3 ネットワーク解析

文献の引用関係および共引用関係は、文献がつくるネットワークとして見る事ができる。我々は各文献セットから引用関係および共引用関係のネットワークを構成し、各種指標の解析を行なう。引用関係のネットワークは、ある文献セットに含まれる文献を頂点、引用を有向（ここでは慣例的に、引用を行なっている文献から引用されている文献への方向）・重みなしの枝とするネットワークである。引用関係は各文献セット内部で完結させた；すなわち、文献セットに含まれない文献を引用していたり、文献セットに含まれない文献から引用されていたりしても、それらの情報は考慮しなかった。引用関係ネットワークでの各頂点（文献）の入次数 in-degree は、その文献を引用しているセット中の文献の数に相当する。引用関係ネットワークでの各頂点（文献）の出次数 out-degree は、その文献が引用しているセット中の文献の数に相当する。引用関係ネットワークでの連結成分 component は、そこに含まれる任意の二頂点間にパスが存在する、極大な部分グラフである。連結成分の検出では、枝の向きは考慮しなかった；すなわち、向きにかかわらず二頂点間に枝が存在すれば、パスに含めてよいこととした。

共引用関係のネットワークは、ある文献セットに含まれる文献を頂点、共引用を無向・重み付き（共引用の回数）の枝とするネットワークである。我々はある文献セットに含まれる文献を引用している文献を、文献セット外のものも含めすべて集め、そこから共引用関係を作成した。共引用関係は各文献セット内部で完結させた；すなわち、文献セットに含まれない文献と同時に引用されていても、その関係は考慮しなかった。共引用関係ネットワークでの各頂点の重みなし次数 degree は、各文献の共引用関係にある文献の数に相当する。重み付き次数 node strength [54] は、各文献の共引用回数の総数に相当する。ネットワーク中の密につながった部分構造は、コミュニティと呼ばれる [10, 11]。今回はネットワークのコミュニティへの分割（各頂点がどのコミュニティに属するかの計算）を、モジュラリティと呼ばれる量を最大化することで行なう [11, 86]。モジュラリティ Q はランダムな分割と比較した分割の有意さを表す量で、

$$Q = \sum_i^C (e_{ii} - (\sum_j^C e_{ij})^2) \quad (6)$$

と定義される。ここで、 C はコミュニティの数、 e_{ij} はコミュニティ i に属する頂点とコミュニティ j に属する頂点をつなぐ枝の重みの和の、全枝の重みの和に対する割合である。 Q は 0 から 1 の間の値を取り、大きいほどそのコミュニティへの分割が有意であることを示す。

モジュラリティの最大化には Louvain 法を用いた [87]。コミュニティ分割結果の評価として、達成されたモジュラリティ Q の値の他に、mixing スコア μ を用いた [88]。これは異なるコミュニティを接続する枝の割合であり、

$$\mu = \frac{1}{W} \sum_i^{M^{\text{ext}}} w_i^{\text{ext}} \quad (7)$$

と定義される。ここで、 M^{ext} は異なるコミュニティ間の枝の総数、 W は全枝の重みの和、 w_i^{ext} は異なるコミュニティ間の枝 i の重みである。この値が小さいほど、コミュニティが明瞭に分割されていることを表す。

3.3 結果

3.3.1 文献セットの基本統計量

前節の手法で構築した文献セットの出版年と被引用数の分布を、それぞれ図 13a, b に示した。太線は DR 文献セットに対応する。細線はそれぞれが 100 個のランダム文献セットのうちの一つに対応する。出版年については、DR 文献セットと比べてランダム文献セットの分布がやや新しい方に寄っていた (図 13a)。また被引用数の分布では、被引用数が少ない部分の分布はほぼ一致しているものの、被引用数が多い部分ではややばらついていた (図 13b)。これら若干の違いはあるものの、意図した通り出版年・被引用数ともに同程度の DR 文献セットとランダム文献セットを構築できた。

文献セットの B スコアの分布を、図 13c に示した。DR 文献セットと比較して、ランダム文献セットでは各文献の B の値は非常に小さかった。文献セットの引用数の分布を、図 13d に示した。特に拘束を設けなかったにもかかわらず、DR 文献セットの方がランダム文献セットよりもやや引用数が少ない傾向が見られた。

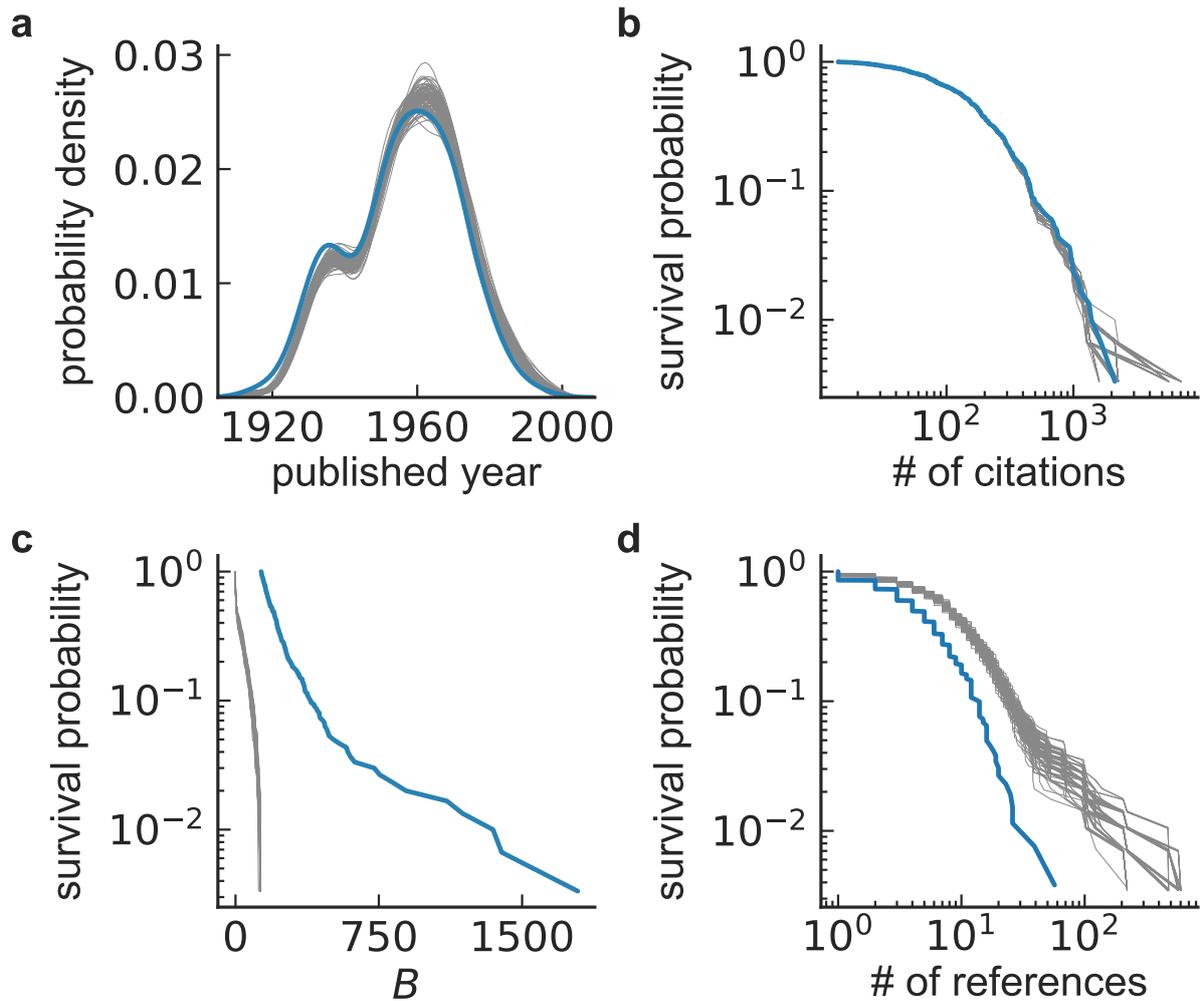


図 13. 文献セットの基本統計量。太線は DR 文献セットに対応する。細線はそれぞれが 100 個のランダム文献セットのうちの一つに対応する。(a) 出版年の分布。(b) 被引用数の survival probability. (c) B スコアの survival probability. (d) 引用数の survival probability.

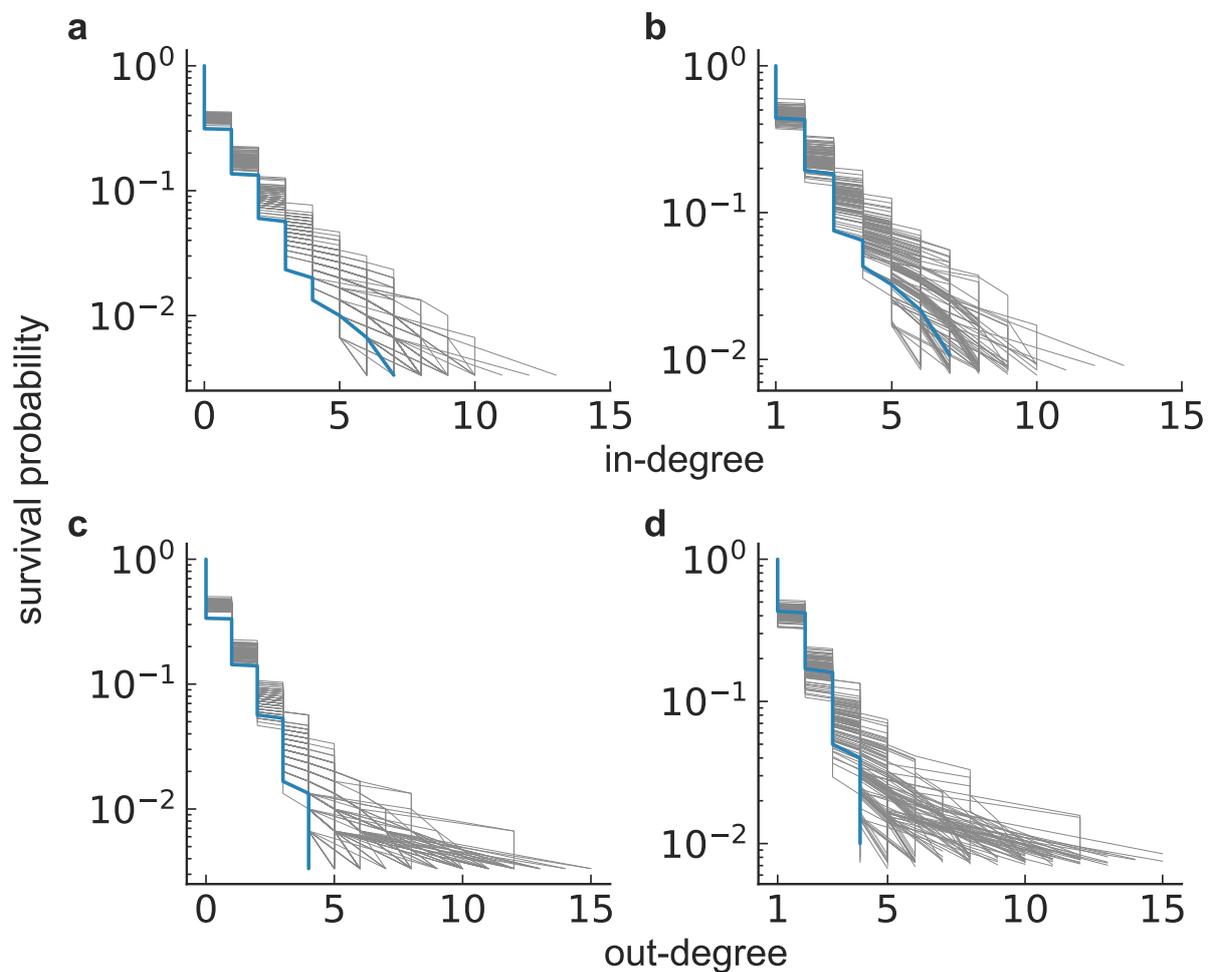


図 14. 引用関係ネットワークでの次数分布。太線は DR 文献セットに対応する。細線はそれぞれが 100 個のランダム文献セットのうちのひとつに対応する。(a) 全頂点の入次数の survival probability. (b) 入次数 = 0 の頂点を除いた入次数の survival probability. (c) 全頂点の出次数の survival probability. (d) 出次数 = 0 の頂点を除いた出次数の survival probability.

3.3.2 引用関係ネットワークの構造

前項で見たとおり、DR 文献に特有の性質を調査するのにおおむね適切な文献セットが構築できたと考えられるため、これら文献セットのつくるネットワークの解析を行なった。引用関係ネットワークでの文献の入次数分布を、図 14a に示した。DR 文献セットでは、ランダム文献セットと比較して、各頂点（文献）の入次数は小さい傾向が見られた。また、入次数 = 0 の頂点も DR 文献セットでは多かった。そこで入次数 = 0 の頂点を除いた場合の傾向を見るため、その survival probability を図 14b に示した。しかしながら、すべての頂点を含めた場合と傾向は変わらなかった。出次数の場合も、傾向は同様であった（図 14c, d）。

引用関係ネットワークでの連結成分の数を、図 15a に示した。DR 文献セットでは、ランダム文献セットと比べて連結成分の数が多かった。DR 文献セットでは次数 = 0 の孤立した頂点も多いが（図 14a, c）、図 15a ではそれら孤立頂点もひとつの連結成分として数えられている。そこでそれらを除いた分布を図 15b に示した。しかし、すべての頂点を含めた場合と傾向は変わらなかった。

各連結成分の密度の分布を、図 15c に示した。DR 文献セットとランダム文献セットとで、各連結成分の密度に顕著な差は見られなかった。各連結成分の頂点数と枝数の関係を、図 15d に示した。ここで、100 個のランダム文献セットは区別せず、すべてのセットでの連結成分をまとめて図示した。密度の分布と一致して、顕著な外れ値は見られなかったものの、DR 文献セットではランダム文献セットと比べて、各連結成分の規模が小さい傾向が見られた。

以上を考慮すると、DR 文献セットがつくる引用関係ネットワークでは、頂点同士が大きな連結成分を形成せず、孤立頂点を含め小さいまとまりに別れているという特徴が考えられる。すなわち、DR 文献セットの引用関係ネットワークでの結びつきは、ランダムと比較しても強くないことが示唆された。

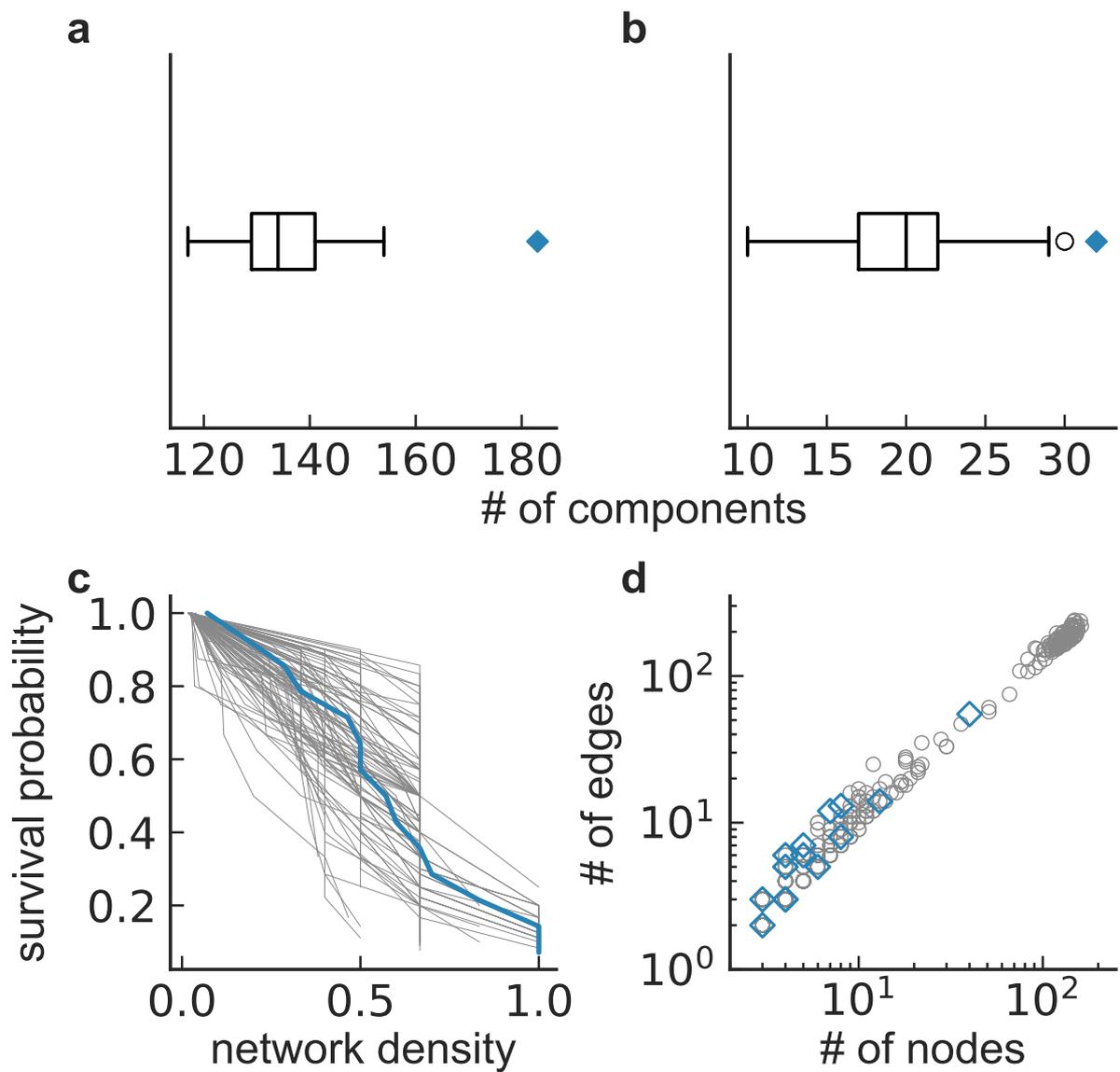


図 15. 引用関係ネットワークでの連結成分に関する各種統計。(a) 連結成分の数。(b) 孤立頂点を除いた連結成分の数。ひし形は DR 文献セットに対応する。箱ひげ図はランダム文献セットに対応する。(c) 連結成分の密度の survival probability. 太線は DR 文献セットに対応する。細線はそれぞれが 100 個のランダム文献セットのうちの一つに対応する。(d) 連結成分の頂点数と枝数の関係。ひし形は DR 文献セットでの各連結成分に対応する。丸はそれぞれがランダム文献セットでの各連結成分に対応する。

表 4. 共引用関係ネットワークにおける最大連結成分の頂点数。最大連結成分に含まれる頂点数と、その値をもつ文献セットの数を示した。

	頂点数	セット数
DR 文献セット	298	1
	295	6
	296	10
ランダム文献セット	297	16
	298	28
	299	21
	300	19

3.3.3 共引用関係ネットワークの構造

DR 文献セットおよびランダム文献セットから作成した共引用関係ネットワークでの、最大連結成分の数を表 4 に示した。どちらの文献セットでも、全頂点 300 個のうちほとんどの頂点がひとつの連結成分を構成した。以下ではそれら最大連結成分（に含まれる頂点）を解析の対象とし、それらに含まれない頂点は解析から除いた。共引用関係ネットワークでの重みなし次数と重み付き次数の分布を、それぞれ図 16a, b に示した。重みなし次数の分布は、DR 文献セットとランダム文献セットとで同程度か、DR 文献セットがやや小さいという傾向が見られた（図 16a）。しかし重み付き次数（共引用の総数）の場合、DR 文献セットはランダム文献セットと比べて大きな次数をもつ傾向が見られた（図 16b）。すなわち DR 文献セットは、枝の重みを考慮すると、共引用関係ネットワークにおいてランダム文献セットよりも強く結びついていることが示唆された。実際に平均共引用回数の分布を見ると、DR 文献セットでランダム文献セットよりも大きかった（図 16c）。

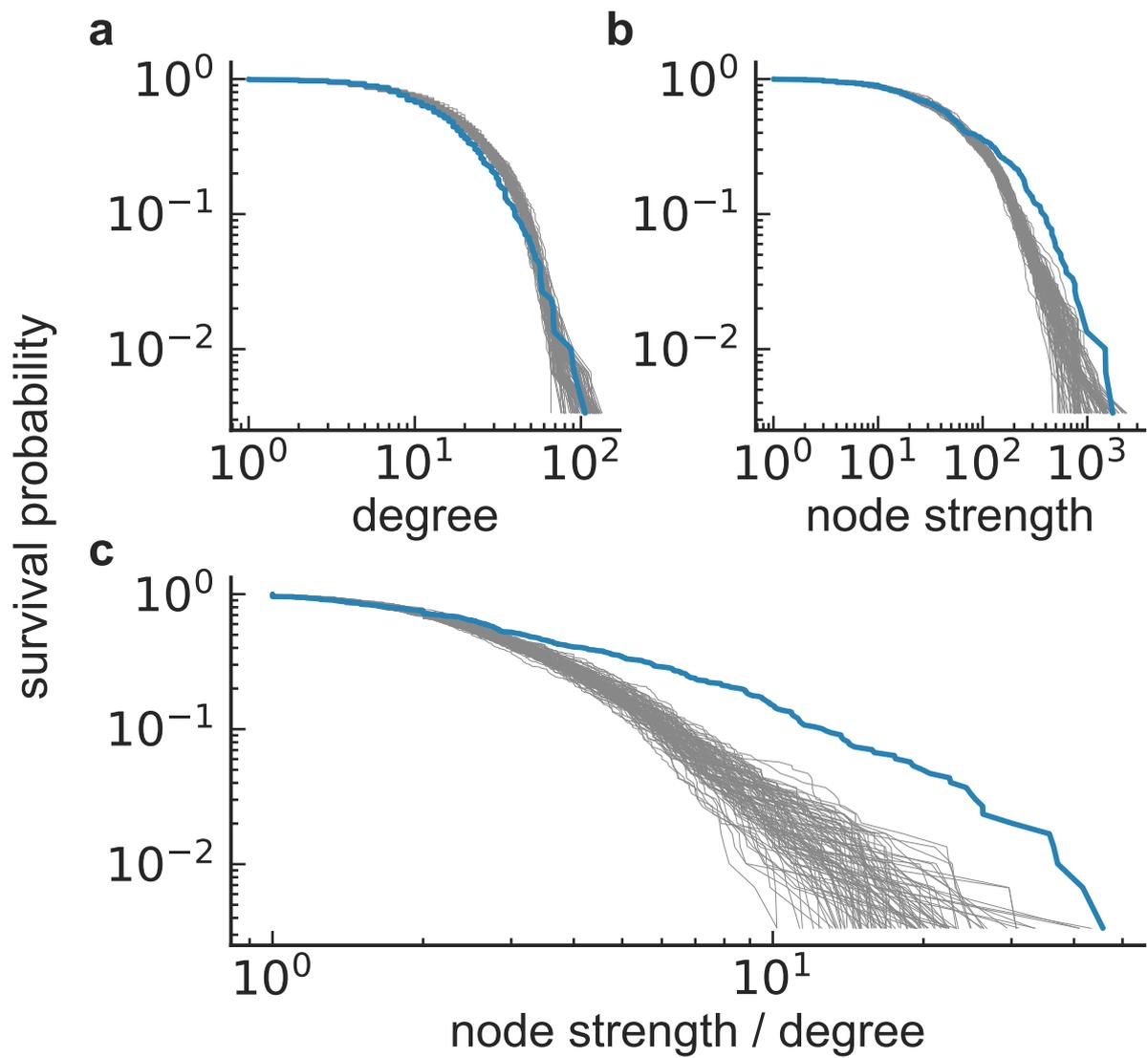


図 16. 共引用関係ネットワークでの度数に関する分布。太線は DR 文献セットに対応する。細線はそれぞれが 100 個のランダム文献セットのうちの一つに対応する。(a) 重みなし次数 degree の survival probability. (b) 重み付き次数 node strength の survival probability. (c) 平均共引用回数 (node strength / degree) の survival probability.

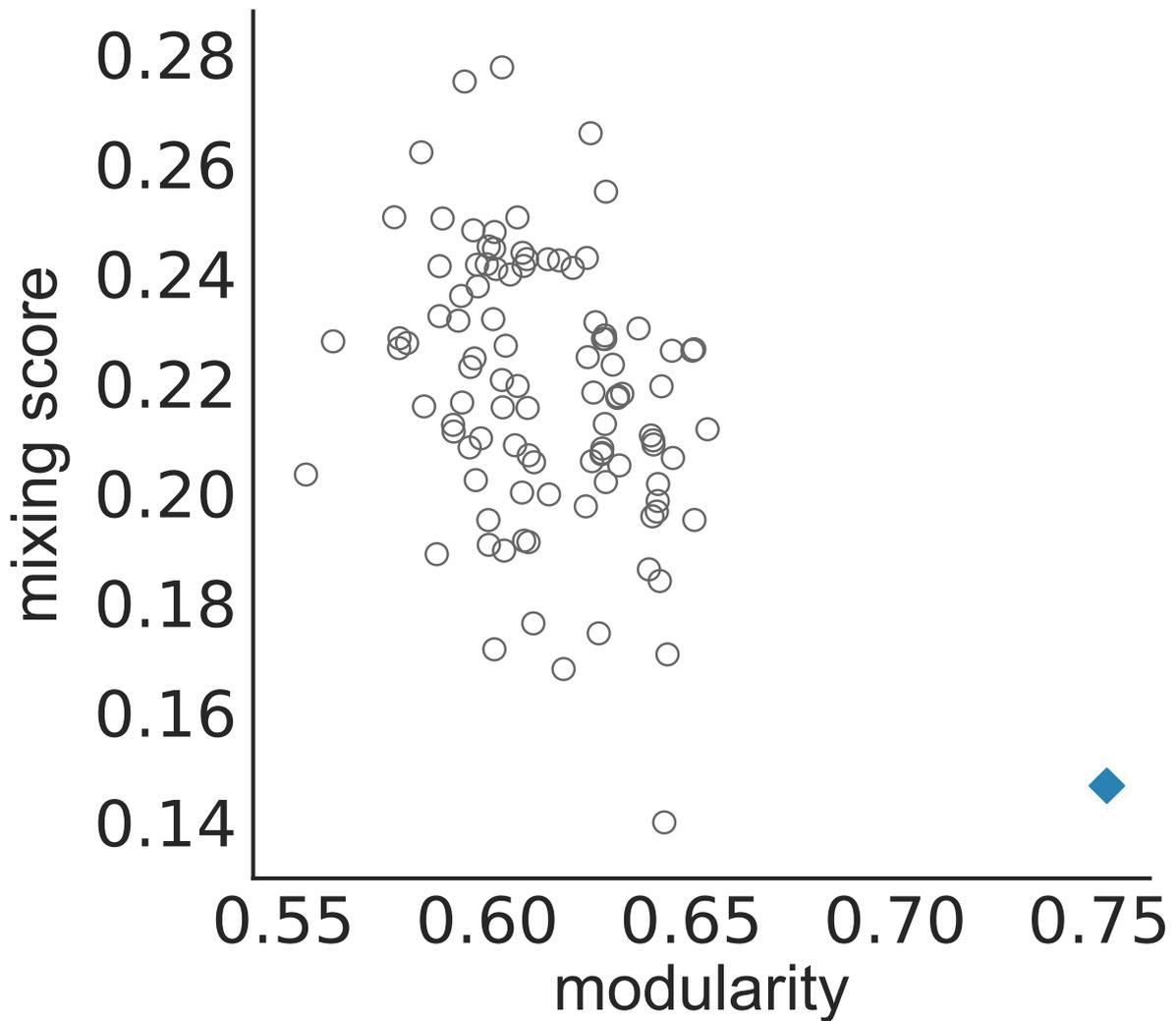


図 17. 共引用関係ネットワークにおけるコミュニティ分割の指標。ひし形は DR 文献セットに対応する。丸はそれぞれが 100 個のランダム文献セットのうちの一つに対応する。

共引用関係ネットワークのコミュニティ分割に関する指標を、図 17 に示した。ランダム文献セットでの結果と比較して、DR 文献セットでの分割では、より高いモジュラリティとより低い mixing スコアを得られた。すなわち、DR 文献セットは共引用関係ネットワークにおいて、ランダム文献セットよりも強いコミュニティを形成していることが示唆された。

表 5. DR 文献の共引用関係ネットワークにおけるコミュニティを構成する文献の例。B スコアの高い順に並べた。DOI は先頭の “10.1103/” を省略した。

DOI	タイトル	出版年	B
PhysRev.82.403	Interaction between the d-Shells in the Transition Metals. II. Ferromagnetic Compounds of Manganese with Perovskite Structure	1951	1790.892
PhysRev.100.675	Considerations on Double Exchange	1955	1347.705
PhysRev.100.545	Neutron Diffraction Study of the Magnetic Properties of the Series of Perovskite-Type Compounds $[(1-x)\text{La}, x\text{Ca}]\text{MnO}_3$	1955	1105.974
PhysRev.118.141	Effects of Double Exchange in Magnetic Crystals	1960	892.117
PhysRev.100.564	Theory of the Role of Covalence in the Perovskite-Type Manganites $[\text{La}, \text{M(II)}]\text{MnO}_3$	1955	819.383
PhysRev.71.809	Finite Elastic Strain of Cubic Crystals	1947	428.107
PhysRev.81.440	Interaction Between the d Shells in the Transition Metals	1951	328.283
PhysRev.79.350	Antiferromagnetism. Theory of Superexchange Interaction	1950	270.868
PhysRev.124.373	Relationship Between Crystal Symmetry and Magnetic Properties of Ionic Compounds Containing Mn^{3+}	1961	157.209
PhysRev.108.1394	Criterion for Ferromagnetism from Observations of Magnetic Isotherms	1957	155.553
PhysRev.45.405	The Dirac Vector Model in Complex Spectra	1934	152.57
PhysRev.155.932	First-Order Localized-Electron \rightleftharpoons Collective-Electron Transition in LaCoO_3	1967	152.297
PhysRevB.5.2382	Coherent-Potential Approximation for a Nonoverlapping-Muffin-Tin-Potential Model of Random Substitutional Alloys	1972	145.431

コミュニティの具体例として、今回のデータ中で最も B スコアの高かった文献を含むコミュニティの内訳を表 5 に示した (図 19j に対応)。また、共引用関係ネットワークにおいてそのコミュニティに属する文献が構成する部分ネットワークを、図 18 に示した。

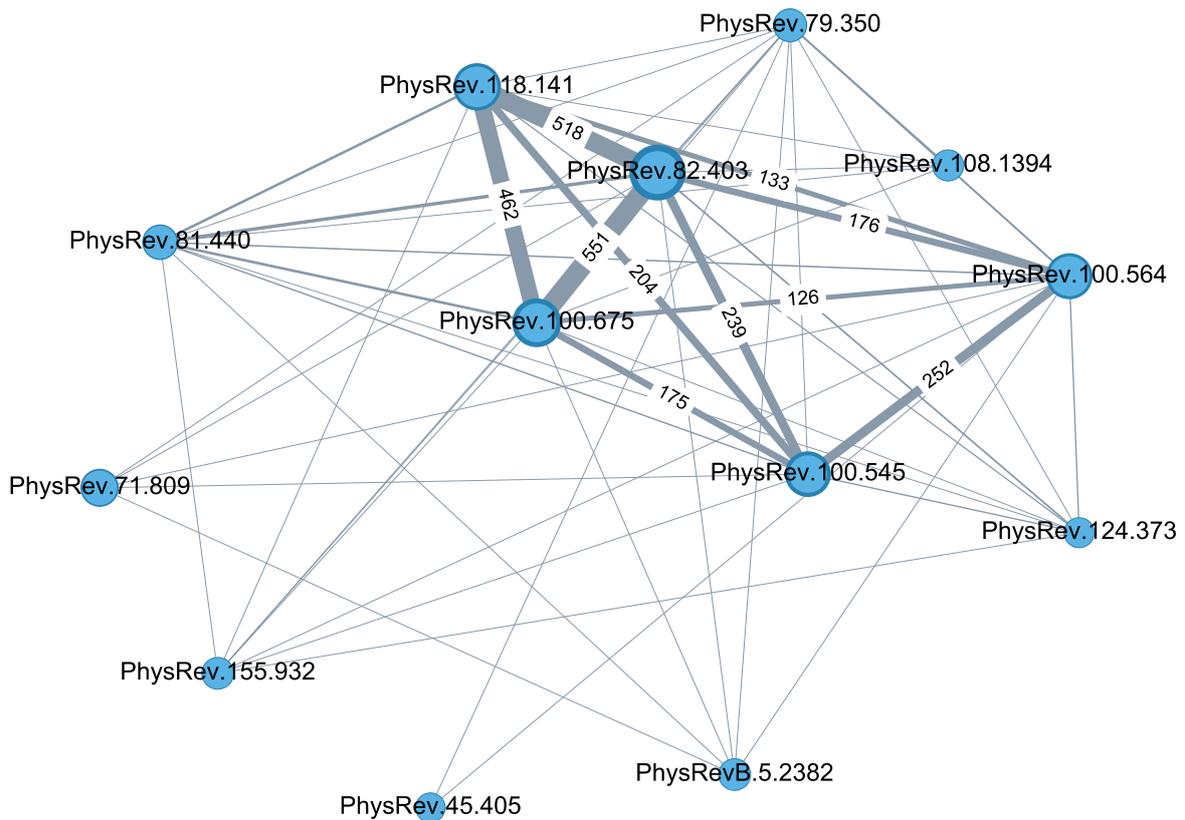


図 18. DR 文献の共引用関係ネットワークにおけるコミュニティの例。頂点は文献、枝は共引用関係を表す。コミュニティ外部と接続する枝は除いた。頂点についたラベルは DOI を表す（先頭の “10.1103/” を省略した）。頂点の大きさは被引用数に対応する。頂点の縁の太さは B スコアに対応する。枝の太さは共引用の回数に対応する。共引用の回数が 100 回以上の枝には、ラベルとして回数を示した。

DR 文献セットに含まれる文献の被引用履歴を、共引用関係ネットワークでのコミュニティ別に図 19 に示した。ここで、被引用数は抽出した文献セットだけでなく、全データから算出した。複数のコミュニティで、被引用履歴の変動が同期している様子が観測された。

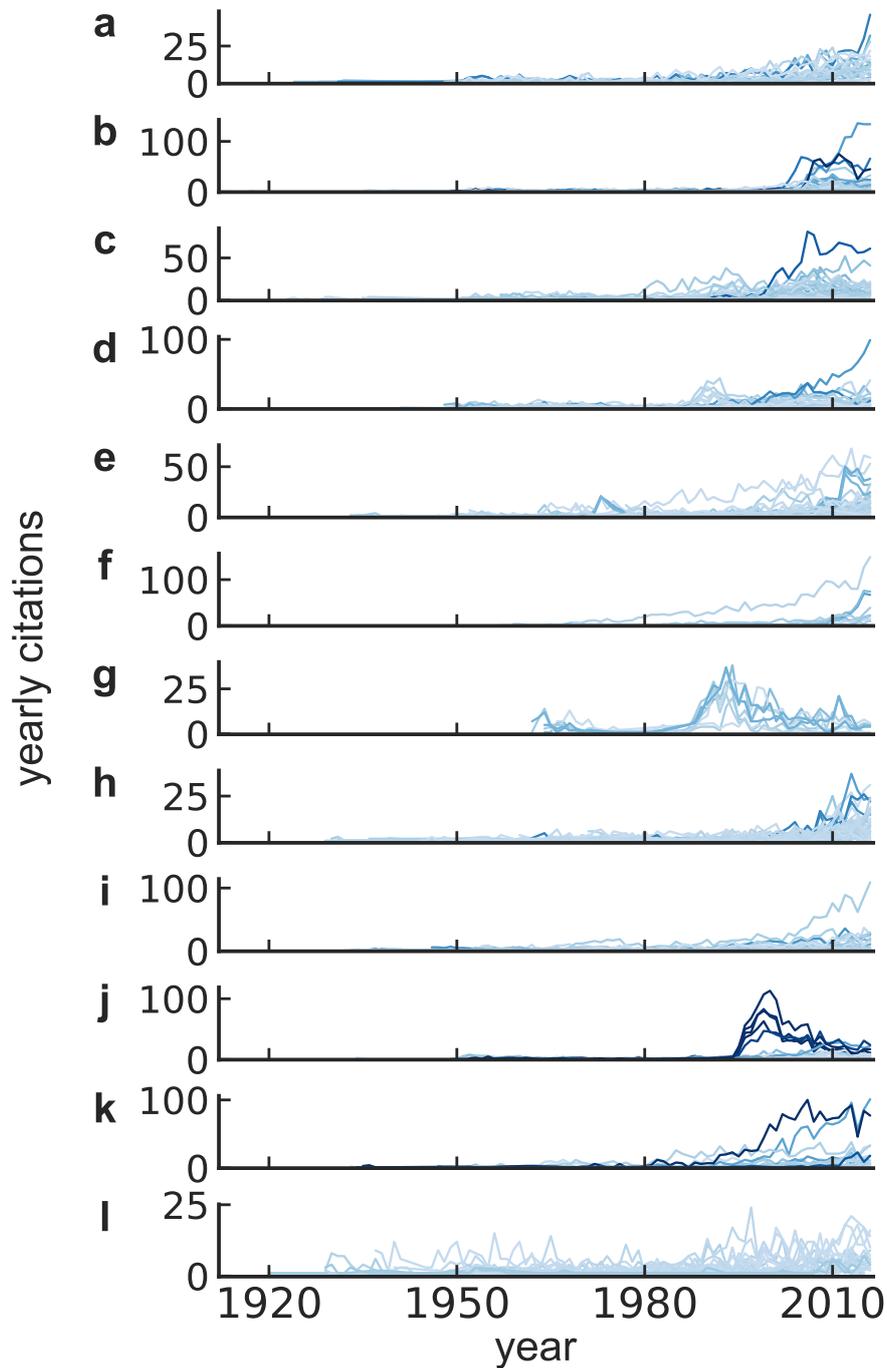


図 19. DR 文献セットの共引用関係ネットワークにおけるコミュニティごとの被引用履歴。線はそれぞれがひとつの文献に対応する。色が濃いほど、その文献の B スコアが高いことを表す。(a) DR 文献 22 本の被引用履歴。(b) DR 文献 26 本の被引用履歴。(c) DR 文献 43 本の被引用履歴。(d) DR 文献 33 本の被引用履歴。(e) DR 文献 44 本の被引用履歴。(f) DR 文献 9 本の被引用履歴。(g) DR 文献 8 本の被引用履歴。(h) DR 文献 39 本の被引用履歴。(i) DR 文献 25 本の被引用履歴。(j) DR 文献 13 本の被引用履歴。(k) DR 文献 16 本の被引用履歴。(l) DR 文献 20 本の被引用履歴。

3.4 考察

以上、我々は DR 文献セットがつくる引用および共引用関係のネットワークの性質を、ランダム文献セットがつくるネットワークとの比較を通して明らかにした。DR 文献セットは引用関係ネットワークにおいては強いつながりをもたないが、共引用関係ネットワークにおいては平均共引用回数、コミュニティ構造などの面で、強いつながりをもつことが示された。また DR 文献セットの共引用関係ネットワークにおけるコミュニティの内訳を見ると、分野ごとのまとまりや、被引用履歴の同期が見られた。

作成した文献セットの統計量を見ると、ランダム文献セットでは DR 文献セットより出版年がやや新しい方に寄っていた (図 13a)。これは DR 文献の定義上、データセット中の比較的古い文献が多いため、より新しい文献の方が数、すなわちランダム抽出の選択肢が多いことによると考えられる。被引用数の分布では、被引用数が少ない部分では DR 文献セットとランダム文献セットでほぼ一致していたが、被引用数が多い部分でばらつきが見られた (図 13b)。これは被引用数が多くなるほど、同じ被引用数をもつ文献が少なくなるためであると考えられる。ランダム文献セットでは B スコアが非常に小さかった (図 13c)。これは、 B の分布はいわゆるべき乗則に従うことが知られており、小さな値をもつ文献が多いためであると考えられる [75]。このことによって、特に条件を課さなくても DR 文献セットとの差を見ることができたが、中程度の B スコアをもつ文献セットの性質を調べることは今後の課題として考えられる。DR 文献セットでは、ランダム文献セットと比べて引用数が少なかった (図 13d)。これは意図しなかったことではあるが、DR 文献セットが先駆的な研究を多く含んでいる可能性があり、それが影響していることが考えられる [75]。

引用関係ネットワークにおいては、DR 文献セットはランダム文献セットと比較して、つながりは同程度あるいは弱めの傾向があった (図 14, 15)。上記の通り文献セットを構築した段階で既に、引用数は DR 文献セットで少ない傾向があった (図 13d)。このことが文献セット内部の引用関係ネットワークに影響していることも考えられる。しかし、出版年と被引用数に加えて引用数も同程度にするような拘束をかけた場合、ランダム文献セットがうまく構築できなかつたので、影響の検証はできなかつた。ただし、これが DR 文献の性質を表しているとも見られる。

共引用関係ネットワークにおいては、DR 文献セットはランダム文献セットと比べて強いつながりを示した (図 16, 17)。引用関係ネットワークとあわせて考えると、DR 文献は、関係する文献が出版された段階で決定する (文献セット内部の) 引用関係ネットワークではつながりが弱い、その後出版された文献によって決まる共引用関係ネットワークではつなが

りが強いということになる。したがって、出版当時にはあまり関連がなかった DR 文献同士の間にも、後の研究によって関連が生まれたと考えることができる。このことが、DR 文献が着目されるきっかけと関係している可能性が考えられる。

共引用関係ネットワークにおけるコミュニティ内部で、被引用履歴の同期が見られた (図 19)。DR 文献のまとまりで被引用履歴が同期するという現象は、既に報告がなされている [75,89]。ただし、本研究ではランダム文献セットとの比較として定量的に DR 文献の関連の強さを示したことで、一定の意義はあると考えられる。また、DR 文献を個別に調べるのではなく、ある程度のまとまりとして調べた方が有意義であることを結果は示唆している。このことをさらに確実に主張できる材料を提供できたことも、成果として考えられる。

4 総括

以上の通り本研究では、オンライン市場での売買関係および文献の引用関係という2つの実世界データを対象として、ネットワークの局所的な構造に着目した解析を行なった。

売買関係ネットワークの解析では、日本最大規模の消費者間電子商取引サービスであるメルカリのデータを解析した。通常ユーザーと不正ユーザーをそれぞれサンプルし、ユーザー同士の取引関係がつくる局所的なネットワーク、egocentric network を作成し、ネットワーク上の指標を解析した。その結果、いくつかの指標で通常ユーザーと不正ユーザーとで、確かに違いが見られた。またそれらの指標から作った特徴量を用いて機械学習を行ない、通常ユーザーと不正ユーザーとを高い確度で分類することができた。このように、利用できる情報がネットワークの局所的な構造に限定されていたにもかかわらず、不正ユーザーの性質を明らかにすることができた。また機械学習を用いた分類器は、大域的な情報を用いた先行研究に匹敵する性能を發揮した。

引用関係ネットワークの解析では、長らく注目されなかったが突然引用されるようになる delayed recognition (DR) 文献を対象とした。DR 文献セットから作成した引用関係および共引用関係ネットワークの特徴を、ランダム文献セットから作成したものと比較することにより、DR 文献に特有の性質を定量的に検討した。その結果、DR 文献セットは共引用関係ネットワークにおいて、ランダム文献セットから作成したネットワークよりも強いつながりを持つことが示された。またこのネットワークでは、コミュニティごとに被引用履歴が同期していることがわかった。このように、DR 文献が多数検出できるということを活かしてサブセットに着目し、工夫した解析を行なうことで、新たな側面から DR 文献の性質を明らかにすることができた。またこれによって、DR 文献の今後の研究に関して方向性を示唆することができた。

導入に記した通り、ネットワークを解析する際には大域的な指標と局所的な指標の両方を用いることで、より深い洞察を得られることが期待される。しかしながら、様々な事情から局所的な構造に着目したい場合があることも記した。本研究では、局所的な構造・指標という制限がある場合でも、工夫した特徴量や、機械学習、複数のネットワークなどを用いることで、データへの一定の理解を得られることを示した。言い換えれば、制限がある中でネットワークを解析するための方法論として、それらの手法を発案し有効性を示すことができた。

この世界から生み出される、我々の行動や生態に関するデータは、今後も増え続けることが容易に予想される。個々の存在と同等かそれ以上に、それらのつながりが重要な意味を持つことも多々あると考えられる。そのような中で、具象の世界と抽象の世界をつなぎ、多様な存在を記述できるネットワークは、ますますその重要性を増していくことが想定される。

ただし、抽象化された世界で大量のデータを扱えるネットワーク解析では、ともすると全体像を美しく図示したものの、具体的に何が読み取れるか、この世界に対して何が言えるかはよくわからない、という状況にもなりがちである。そういった陥穽が考えられる中で、局所構造、あるいは小さなネットワークに注目することが、逆説的に重要になってくることを推し量ることもできる。つまり、大量のデータ全体をうまく利用する手法とともに、細かな構造・特徴の差異に着目することについても、洗練された手法が望まれるはずである。そのような観点から、本研究は一定の貢献をすることができたと考える。

謝辞

木下賢吾教授には、学部4年時の研究室配属より今に至るまで研究の指導をしていただきました。木下・大林・西研究室の皆様には、日々議論を交わしたり食事をともにしたり、またつらい時に励ましを頂いたりとお世話になりました。データ科学国際共同大学院プログラムにより留学した際には、増田直紀先生に指導していただきました。増田研の皆様との日々によって、慣れない土地での生活は楽しく、また実り多いものとなりました。また東北大学漫画研究会の皆様には、サークルでの活動・交流がなければ今の私はなかったと言えるほどの影響を受けました。最後に、両親には、これまでの気ままな人生を支えていただきました。休学を含めた幾度かの難局を越してここまで来ることができたのは、周りの人々の存在あってのものです。上に挙げた皆様、またここには挙げられなかったけれどもこれまで関わったすべての皆様に感謝したいと思います。

参考文献

1. David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009.
2. Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005.
3. Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.
4. Mark EJ Newman. *Networks: An introduction*. Oxford University Press, Oxford, 2010.
5. Duncan J Watts and Steven H Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684):440, 1998.
6. Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
7. Derek J De Solla Price. Networks of scientific papers. *Science*, pages 510–515, 1965.
8. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30:107–117, 1998.
9. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.
10. Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.
11. Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.
12. Stephen P Borgatti, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, 2009.
13. Albert-László Barabási and Zoltan N Oltvai. Network biology: understanding the

- cell’s functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.
14. David Mears and Harvey B Pollard. Network science and the human brain: using graph theory to understand the brain and one of its hubs, the amygdala, in health and disease. *Journal of neuroscience research*, 94(6):590–605, 2016.
 15. UK Parliament. The growing threat of online fraud. <https://www.parliament.uk/business/committees/committees-a-z/commons-select/public-accounts-committee/inquiries/parliament-2017/growing-threat-online-fraud-17-19/publications/>, 2017 (Accessed: April 25, 2019).
 16. McAfee, LLC. Economic impact of cybercrime report. <https://www.mcafee.com/enterprise/en-us/solutions/lp/economics-cybercrime.html>, 2018 (Accessed: April 25, 2019).
 17. Ross Anderson, Chris Barton, Rainer Böhme, Richard Clayton, Michel JG Van Eeten, Michael Levi, Tyler Moore, and Stefan Savage. Measuring the cost of cybercrime. In *The economics of information security and privacy*, pages 265–300. Springer, Berlin, Heidelberg, 2013.
 18. Richard J Bolton and David J Hand. Statistical fraud detection: A review. *Statistical Science*, 17:235–249, 2002.
 19. Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. Preprint arXiv:1009.6119, 2010.
 20. Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal. Fraud detection system: A survey. *J. Netw. Comput. Appl.*, 68:90–113, 2016.
 21. Jarrod West and Maumita Bhattacharya. Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, 57:47–66, 2016.
 22. Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. *Data Min. Knowl. Discov.*, 29:626–688, 2015.
 23. Google LLC and White Ops, Inc. The hunt for 3ve. https://services.google.com/fh/files/blogs/3ve_google_whiteops_whitepaper_final_nov_2018.pdf, 2018 (Accessed: May 10, 2019).
 24. Calton Pu and Steve Webb. Observed trends in spam construction techniques: A case study of spam evolution. In *CEAS*, pages 104–112, 2006.
 25. Brian Hayes. How many ways can you spell v1@gra? *Amer. Sci.*, 95:298–302, 2007.

26. Alexy Bhowmick and Shyamanta M Hazarika. Machine learning for e-mail spam filtering: review, techniques and trends. Preprint arXiv:1606.01042, 2016.
27. David Savage, Xiuzhen Zhang, Xinghuo Yu, Pauline Chou, and Qingmai Wang. Anomaly detection in online social networks. *Soc. Netw.*, 39:62–70, 2014.
28. Lovro Šubelj, Štefan Furlan, and Marko Bajec. An expert system for detecting automobile insurance fraud using social network analysis. *Expert Syst. Appl.*, 38:1039–1052, 2011.
29. Rafał Dreżewski, Jan Sepielak, and Wojciech Filipkowski. The application of social network analysis algorithms in a system supporting money laundering detection. *Info. Sci.*, 295:18–32, 2015.
30. Andrea Fronzetti Colladon and Elisa Remondi. Using social network analysis to prevent money laundering. *Expert Syst. Appl.*, 67:49–58, 2017.
31. David Savage, Qingmai Wang, Xiuzhen Zhang, Pauline Chou, and Xinghuo Yu. Detection of money laundering groups: Supervised learning on small networks. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, pages 43–49, 2017.
32. Juan Liu, Eric Bier, Aaron Wilson, John Alexis Guerra-Gomez, Tomonori Honda, Kumar Sricharan, Leilani Gilpin, and Daniel Davies. Graph analysis for detecting fraud, waste, and abuse in healthcare data. *AI Magazine*, 37:33–46, 2016.
33. Oleksandr Shchur, Aleksandar Bojchevski, Mohamed Farghal, Stephan Günnemann, and Yusuf Saber. Anomaly detection in car-booking graphs. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 604–607, 2018.
34. Véronique Van Vlasselaer, Tina Eliassi-Rad, Leman Akoglu, Monique Snoeck, and Bart Baesens. Gotcha! network-based fraud detection for social security fraud. *Manag. Sci.*, 63:3090–3110, 2016.
35. Jinlong Hu, Junjie Liang, and Shoubin Dong. ibgp: A bipartite graph propagation approach for mobile advertising fraud detection. *Mobile Info. Syst.*, 2017:1–12, 2017.
36. Emilio Ferrara, Pasquale De Meo, Salvatore Catanese, and Giacomo Fiumara. Detecting criminal organizations in mobile phone networks. *Expert Syst. Appl.*, 41:5733–5750, 2014.

37. Sajid Yousuf Bhat and Muhammad Abulaish. Community-based features for identifying spammers in online social networks. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 100–107, 2013.
38. Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. Inferring strange behavior from connectivity pattern in social networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 126–138, 2014.
39. Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. Fraudar: Bounding graph fraud in the face of camouflage. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 895–904, 2016.
40. Jawairia Rasheed, Usman Akram, and Ahmed Kamran Malik. Terrorist network analysis and identification of main actors using machine learning techniques. In *Proceedings of the 6th International Conference on Information Technology: IoT and Smart City*, pages 7–12, 2018.
41. Leman Akoglu, Rishi Chandy, and Christos Faloutsos. Opinion fraud detection in online reviews by network effects. In *Seventh International AAAI Conference on Weblogs and Social Media*, pages 2–11, 2013.
42. Shenghua Liu, Bryan Hooi, and Christos Faloutsos. Holoscope: Topology-and-spike aware fraud detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1539–1548, 2017.
43. Zhuo Wang, Songmin Gu, Xiangnan Zhao, and Xiaowei Xu. Graph-based review spammer group detection. *Knowl. Info. Syst.*, 55:571–597, 2018.
44. Duen Horng Chau, Shashank Pandit, and Christos Faloutsos. Detecting fraudulent personalities in networks of online auctioneers. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 103–114, 2006.
45. Shashank Pandit, Duen Horng Chau, Samuel Wang, and Christos Faloutsos. Net-probe: A fast and scalable system for fraud detection in online auction networks. In *Proceedings of the 16th International Conference on World Wide Web*, pages 201–210, 2007.
46. Jyun-Cheng Wang and Chui-Chen Chiu. Recommending trusted online auction sellers using social network analysis. *Expert Syst. Appl.*, 34:1666–1679, 2008.

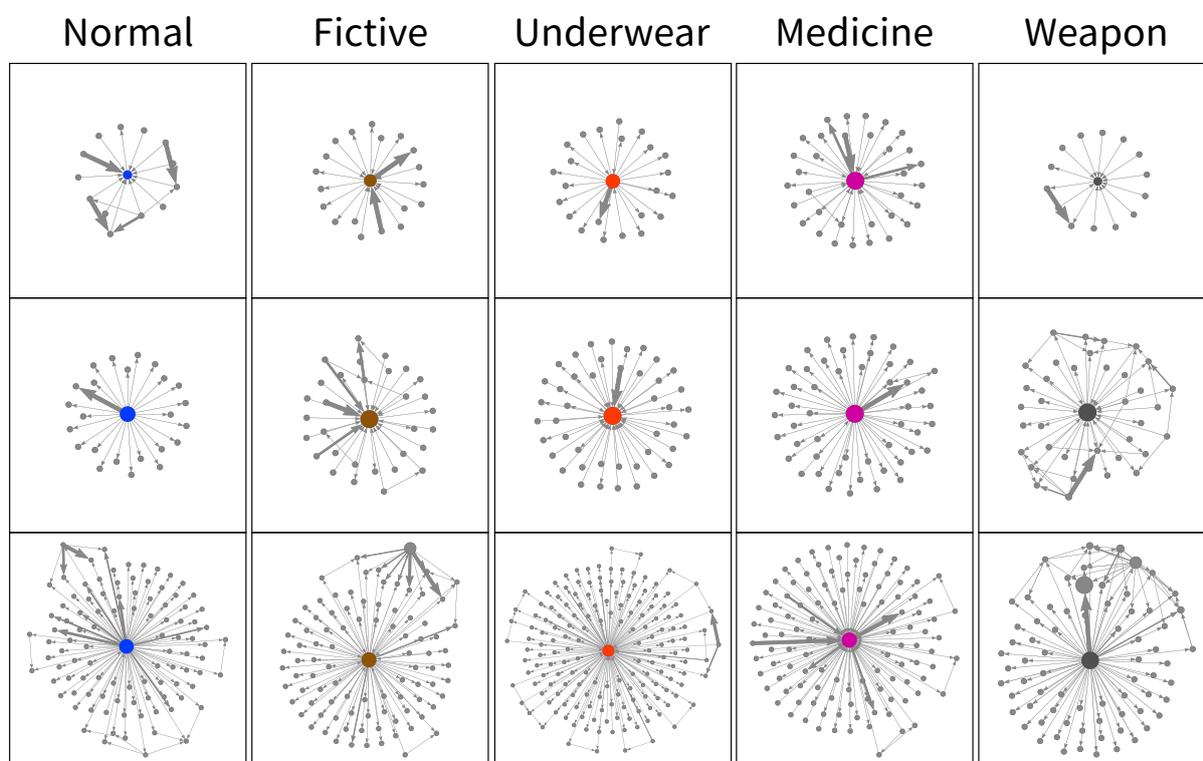
47. Phiradet Bangcharoensap, Hayato Kobayashi, Nobuyuki Shimizu, Satoshi Yamauchi, and Tsuyoshi Murata. Two step graph-based semi-supervised learning for online auction fraud detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 165–179, 2015.
48. Zhu Yanchun, Zhang Wei, and Yu Changhai. Detection of feedback reputation fraud in taobao using social network theory. In *2011 International Joint Conference on Service Sciences*, pages 188–192, 2011.
49. Véronique Van Vlasselaer, Cristián Bravo, Olivier Caelen, Tina Eliassi-Rad, Leman Akoglu, Monique Snoeck, and Bart Baesens. Apate: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Syst.*, 75:38–48, 2015.
50. Yuan Li, Yiheng Sun, and Noshir Contractor. Graph mining assisted semi-supervised learning for fraudulent cash-out detection. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 546–553, 2017.
51. Patrick Monamo, Vukosi Marivate, and Bheki Twala. Unsupervised learning for robust bitcoin fraud detection. In *2016 Information Security for South Africa (ISSA)*, pages 129–134, 2016.
52. Leman Akoglu, Mary McGlohon, and Christos Faloutsos. Oddball: Spotting anomalies in weighted graphs. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 410–421, 2010.
53. Mercari, Inc. Fast facts. <https://about.mercari.com/press/fast-fact/>, (Accessed: April 25, 2019).
54. Alain Barrat, Marc Barthélemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA*, 101:3747–3752, 2004.
55. Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. USA*, 101:2658–2663, 2004.
56. Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, 2005.

57. Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827, 2002.
58. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *J. Machine Learning Res.*, 12:2825–2830, 2011.
59. Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
60. Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and Regression Trees*. Chapman & Hall/CRC, Boca Raton, FL, 1984.
61. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York, NY, 2009.
62. Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:25, 2007.
63. André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: A corrected feature importance measure. *Bioinformatics*, 26:1340–1347, 2010.
64. Hikaru Yamamoto, Nina Sugiyama, Fujio Toriumi, Hikaru Kashida, and Takuma Yamaguchi. Angels or demons? classifying desirable heavy users and undesirable power sellers in online c2c marketplace. *SSRN preprint*, 2018.
65. James A Evans and Jacob G Foster. Metaknowledge. *Science*, 331(6018):721–725, 2011.
66. Richard Van Noorden, Brendan Maher, and Regina Nuzzo. The top 100 papers. *Nature News*, 514(7524):550, 2014.
67. Mark EJ Newman. The first-mover advantage in scientific publication. *EPL (Europhysics Letters)*, 86(6):68001, 2009.
68. Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.
69. Bernard Barber. Resistance by scientists to scientific discovery. *Science*, 134(3479):596–602, 1961.
70. Stephen Cole. Professional standing and the reception of scientific discoveries.

- American Journal of Sociology*, 76(2):286–306, 1970.
71. Eugene Garfield. Premature discovery or delayed recognition-why? *Current Contents*, (21):5–10, 1980.
 72. Anthony FJ van Raan. Sleeping beauties in science. *Scientometrics*, 59(3):467–472, 2004.
 73. Ronald Rousseau. Delayed recognition: recent developments and a proposal to study this phenomenon as a fuzzy concept. *Journal of Data and Information Science*, 3(3):1–13, 2018.
 74. Cassidy R Sugimoto and Javed Mostafa. A note of concern and context: On careful use of terminologies. *Journal of the Association for Information Science and Technology*, 69(3):347–348, 2018.
 75. Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, 112(24):7426–7431, 2015.
 76. Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4):265–269, 1973.
 77. Tibor Braun, Wolfgang Glänzel, and András Schubert. On sleeping beauties, princes and other tales of citation distributions... *Research Evaluation*, 19(3):195–202, 2010.
 78. Norio Ohba and Kumiko Nakao. Sleeping beauties in ophthalmology. *Scientometrics*, 93(2):253–264, 2012.
 79. Jiang Li. Citation curves of “all-elements-sleeping-beauties” : “flash in the pan” first and then “delayed recognition” . *Scientometrics*, 100(2):595–601, 2014.
 80. Werner Marx. The shockley-queisser paper—a notable example of a scientific sleeping beauty. *Annalen der Physik*, 526(5-6):A41–A45, 2014.
 81. You Song, Fangling Situ, Hongjun Zhu, and Jinzhi Lei. To be the prince to wake up sleeping beauty: the rediscovery of the delayed recognition studies. *Scientometrics*, 117(1):9–24, 2018.
 82. Jian Du and Yishan Wu. A bibliometric framework for identifying “princes” who wake up the “sleeping beauty” in challenge-type scientific discoveries. *Journal of Data and Information Science*, 1(1):50–68, 2016.

83. Aurora AC Teixeira, Pedro Cosme Vieira, and Ana Patrícia Abreu. Sleeping beauties and their princes in innovation studies. *Scientometrics*, 110(2):541–580, 2017.
84. Adil El Aichouchi and Philippe Gorry. Delayed recognition of judah folkman’ s hypothesis on tumor angiogenesis: when a prince awakens a sleeping beauty by self-citation. *Scientometrics*, 116(1):385–399, 2018.
85. Hui Fang. A transition stage co-citation criterion for identifying the awakeners of sleeping beauty publications. *Scientometrics*, 121(1):307–322, 2019.
86. Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.
87. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
88. Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110, 2008.
89. Sidney Redner. Citation statistics from 110 years of physical review. *Physics today*, 58(6):49–54, 2005.

付録



図付録.1. 売買関係：各種ユーザーの ego-centric network の例。

表付録.1. DR 文献の共引用関係ネットワークにおけるコミュニティとそれを構成する文献。

コミュニティ	タイトル	出版年	<i>B</i>
a	Reciprocal Relations in Irreversible Processes. I.	1931	601.714
a	Reciprocal Relations in Irreversible Processes. II.	1931	424.255
a	Thermal Agitation of Electric Charge in Conductors	1928	381.317
a	The Dynamics of Capillary Flow	1921	359.089
a	On the Resistance Experienced by Spheres in their Motion through Gases	1924	349.009
a	Thermal Agitation of Electricity in Conductors	1928	322.519
a	On the Theory of the Brownian Motion	1930	320.381
a	Theory of Thermal Transport Coefficients	1964	261.673
a	Three-Level Masers as Heat Engines	1959	235.596
a	Theory of Radiative Heat Transfer between Closely Spaced Bodies	1971	219.832
a	The Thermodynamics of Irreversible Processes. III. Relativistic Theory of the Simple Fluid	1940	219.394
a	A Model for Collision Processes in Gases. I. Small Amplitude Processes in Charged and Neutral One-Component Systems	1954	218.357
a	On Onsager's Principle of Microscopic Reversibility	1945	216.524
a	Observation of Anderson Localization in an Electron Gas	1969	190.477
a	Surface Plasmons in Thin Films	1969	190.159
a	Diffraction Theory of Electromagnetic Waves	1939	178.066
a	Thermodynamics and Statistical Mechanics at Negative Absolute Temperatures	1956	176.37
a	Network theory of microscopic and macroscopic behavior of master equation systems	1976	167.662
a	Fluctuations and Irreversible Processes	1953	157.1
a	A Nuclear Spin System at Negative Temperature	1951	151.754
a	Single-particle distribution in the hydrodynamic and statistical thermodynamic models of multiparticle production	1974	135.737
a	Irreversibility and Generalized Noise	1951	133.977

コミュニティ	タイトル	出版年	B
b	The Band Theory of Graphite	1947	1188.701
b	Spin-Orbit Coupling Effects in Zinc Blende Structures	1955	624.16
b	Hall Effect in Ferromagnetics	1954	483.272
b	Model for a Quantum Hall Effect without Landau Levels: Condensed-Matter Realization of the "Parity Anomaly"	1988	477.597
b	Diamagnetism of Graphite	1956	413.603
b	Band Structure of Graphite	1958	410.954
b	Accidental Degeneracy in the Energy Bands of Crystals	1937	394.434
b	Lattice Constants of Graphite at Low Temperatures	1955	269.275
b	Condensed-Matter Simulation of a Three-Dimensional Anomaly	1984	253.133
b	Mechanical Detection and Measurement of the Angular Mo- mentum of Light	1936	224.176
b	Self-consistent effective-mass theory for intralayer screening in graphite intercalation compounds	1984	212.045
b	Band Structure of Graphite and de Haas-van Alphen Effect	1957	207.136
b	Side-Jump Mechanism for the Hall Effect of Ferromagnets	1970	206.494
b	On the Surface States Associated with a Periodic Potential	1939	192.522
b	Cyclotron Resonance in Graphite	1958	180.745
b	Calculation and Experimental Proof of the Transverse Shift Induced by Total Internal Reflection of a Circularly Polarized Light Beam	1972	174.14
b	Two applications of axion electrodynamics	1987	170.486
b	Magnetization by Rotation	1915	167.531
b	Theory of the Hall Effect in Ferromagnetic Substances	1958	164.108
b	Theory of the Effect of Spin-Orbit Coupling on Magnetic Res- onance in Some Semiconductors	1954	163.478
b	Calculation of the g Factor of Metallic Sodium	1952	157.616
b	Vortices and Streams Caused by Sound Waves	1948	144.374
b	The Structure of Electronic Excitation Levels in Insulating Crystals	1937	140.327
b	Band Structure and Laser Action in $Pb_xSn_{1-x}Te$	1966	138.277
b	The Widths of the L-Series Lines and of the Energy Levels of Au(79)	1934	136.543
b	A remnant of chiral symmetry on the lattice	1982	135.806

コミュニティ	タイトル	出版年	B
c	Superconductivity in a Strong Spin-Exchange Field	1964	725.898
c	The Complete Equation of State of One, Two and Three-Dimensional Gases of Hard Elastic Spheres	1936	457.637
c	Upper Limit for the Critical Field in Hard Superconductors	1962	360.548
c	Possible Pairing without Superconductivity at Low Carrier Concentrations in Bulk and Thin-Film Superconducting Semiconductors	1969	356.359
c	Exact Analysis of an Interacting Bose Gas. I. The General Solution and the Ground State	1963	315.551
c	Quantum-Mechanical Many-Body Problem with Hard-Sphere Interaction	1957	297.14
c	Theory of the Superfluidity of Helium II	1941	296.523
c	Can a Solid Be "Superfluid"?	1970	278.13
c	Theory of the Insulating State	1964	265.176
c	The Electrical Properties of Black Phosphorus	1953	261.803
c	On the Interaction of Electrons in Metals	1934	257.661
c	Many-Body Problem in Quantum Mechanics and Quantum Statistical Mechanics	1957	252.381
c	Speculations on Bose-Einstein Condensation and Quantum Crystals	1970	250.607
c	Anisotropy of the Electronic Work Function of Metals	1941	238.433
c	Bardeen-Cooper-Schrieffer Theory of Superconductivity in the Case of Overlapping Bands	1959	233.199
c	Exact Analysis of an Interacting Bose Gas. II. The Excitation Spectrum	1963	226.484
c	Bose-Einstein Condensation and Liquid Helium	1956	226.265
c	Magnetic and Semiconducting Properties of SmB6	1969	215.297
c	Note on an Approximation Treatment for Many-Electron Systems	1934	210.899
c	Microscopic Theory of Superconductivity	1957	208.192
c	A Collective Description of Electron Interactions: II. Collective vs Individual Particle Aspects of the Interactions	1952	206.447
c	Generalized Diffusion Equation for Superconducting Alloys	1970	203.22
c	Excitonic Insulator	1967	201.678
c	Theory of Electrical Double Layers in Adsorbed Films	1935	199.032
c	Quantum Cell Model for Bosons	1963	195.077
c	Forces in Molecules	1939	192.884
c	The Interaction Between a Neutron and a Proton and the Structure of H3	1935	188.698
c	Further Considerations on Electromagnetic Potentials in the Quantum Theory	1961	166.618

c	Exact Results for a Quantum Many-Body Problem in One Dimension. II	1972	166.295
c	Dislocation Models of Crystal Grain Boundaries	1950	163.531
c	Ground-State Energy of a Many-Fermion System. II	1960	161.767
c	Some Theorems on the Free Energies of Crystal Surfaces	1951	156.496
c	Stability of Spin-Aligned Hydrogen at Low Temperatures and High Magnetic Fields: New Field-Dependent Scattering Resonances and Predissociations	1976	156.345
c	Thermionic Phenomena Caused by Vapors of Rubidium and Potassium	1926	156.105
c	Metallic Hydrogen: A High-Temperature Superconductor?	1968	149.023
c	Compton Modified Line Structure and its Relation to the Electron Theory of Solid Bodies	1929	147.333
c	Electron Mobility in Semiconducting Strontium Titanate	1967	146.998
c	Bose-Einstein Condensation of Excitons	1962	144.286
c	The Effect of Space Charge and Initial Velocities on the Potential Distribution and Thermionic Current between Parallel Plane Electrodes	1923	142.881
c	Theoretical Considerations Concerning Quantized Magnetic Flux in Superconducting Cylinders	1961	142.815
c	Exact Results for a Quantum Many-Body Problem in One Dimension	1971	137.494
c	Size Effect in a Granular Superconductor	1968	136.174
c	Electronic Transport in Semimetallic Cerium Sulfide	1964	134.595

コミュニティ	タイトル	出版年	B
d	New Magnetic Anisotropy	1956	580.473
d	Anisotropic Superexchange Interaction and Weak Ferromagnetism	1960	489.316
d	New Magnetic Anisotropy	1957	376.665
d	On the Theory of Ferromagnetic Resonance Absorption	1948	349.629
d	Thermal Fluctuations of a Single-Domain Particle	1963	338.52
d	Prebreakdown Current in the Townsend Discharge	1956	255.092
d	New Mechanism of Anisotropic Superexchange Interaction	1960	237.777
d	Steady-State Motion of Magnetic Domains	1973	237.301
d	Surface Studies of Solids by Total Reflection of X-Rays	1954	235.37
d	Neutron Diffraction by Paramagnetic and Antiferromagnetic Substances	1951	218.804
d	Theory of Dipole Interaction in Crystals	1946	215.605
d	An Approximate Quantum Theory of the Antiferromagnetic Ground State	1952	211.219
d	Ferromagnetism in a Narrow, Almost Half-Filled s Band	1966	206.23
d	Magnetic and Crystallographic Properties of Substituted Yttrium-Iron Garnet, $3Y_2O_3 \cdot xM_2O_3 \cdot (5-x)Fe_2O_3$	1958	184.484
d	Theory of the Structure of Ferromagnetic Domains in Films and Small Particles	1946	178.315
d	Theory of Spin-Wave Interactions in Ferro- and Antiferromagnetism	1960	176.306
d	Anisotropy of the Magnetoelectric Effect in Cr_2O_3	1961	175.717
d	Field Dependence of the Intrinsic Domain Magnetization of a Ferromagnet	1940	172.101
d	Physical Theory of Ferromagnetic Domains	1949	172.064
d	Antiferromagnetism. The Triangular Ising Net	1950	170.627
d	Ordering and Antiferromagnetism in Ferrites	1956	166.926
d	Role of Anisotropic Exchange Interactions in Determining the Properties of Spin-Glasses	1980	165.92
d	The Spin-Wave Theory of Antiferromagnetics	1952	164.362
d	Ferromagnetic Resonance in Nickel Ferrite Between One and Two Kilomegacycles	1955	156.621
d	Single-Particle Excitations in Magnetic Insulators	1970	149.773
d	Effect of magnon-phonon thermal relaxation on heat transport by magnons	1977	147.192
d	Interaction of Spin Waves and Ultrasonic Waves in Ferromagnetic Crystals	1958	145.067
d	Magnetostatic Modes in Ferromagnetic Resonance	1957	141.619
d	Excitation of Spin Waves in a Ferromagnet by a Uniform rf Field	1958	139.181

d	Upper Bound on the Magnetoelectric Susceptibility	1968	138.233
d	Theory of Ferromagnetism and the Ordering of Electronic Energy Levels	1962	136.139
d	Dielectric Properties of Lead Zirconate	1951	135.698
d	Polarization Effects in the Magnetic Elastic Scattering of Slow Neutrons	1963	133.979

コミュニティ	タイトル	出版年	<i>B</i>
e	Broken Symmetries and the Masses of Gauge Bosons	1964	393.044
e	Global Conservation Laws and Massless Particles	1964	363.042
e	Broken Symmetry and the Mass of Gauge Vector Mesons	1964	350.404
e	Equilibrium parity-violating current in a magnetic field	1980	334.962
e	Selection Rules for the Dematerialization of a Particle into Two Photons	1950	278.734
e	Elementary Particle Theory of Composite Particles	1963	267.671
e	Collision of Two Light Quanta	1934	263.472
e	The Positive Electron	1933	262.747
e	Nature of the High Energy Particles of Penetrating Radiation and Status of Ionization and Radiation Formulae	1934	254.621
e	End to the Cosmic-Ray Spectrum?	1966	251.671
e	On Transition Probabilities in Double Beta-Disintegration	1939	225.487
e	Evidence That the Deuteron Is Not an Elementary Particle	1965	224.083
e	Neutrino masses, mixings, and oscillations in $SU(2) \times U(1)$ models of electroweak interactions	1980	221.453
e	Majorana Neutrinos and the Production of the Right-Handed Charged Gauge Boson	1983	215.663
e	An Identity in Quantum Electrodynamics	1950	214.878
e	Plasmons, Gauge Invariance, and Mass	1963	208.85
e	Evidence for the Formation of Positronium in Gases	1951	204.06
e	Diagonal neutral currents	1977	202.168
e	Erratum: Lepton number as the fourth "color"	1975	201.425
e	Resonant Scattering of Antineutrinos	1960	193.01
e	Bremsstrahlung and Pair Production in Condensed Media at High Energies	1956	191.589
e	Spontaneously Growing Transverse Waves in a Plasma Due to an Anisotropic Velocity Distribution	1959	190.582
e	Molecular Charmonium: A New Spectroscopy?	1977	184.674
e	Symmetry breaking and scalar bosons	1976	181.577
e	Model Calculation for the $Y_0^*(1405)$ Resonance State	1967	180.533
e	Lorentz Noninvariance of the Complex-Ghost Relativistic Field Theory	1971	177.0
e	Pattern of symmetry breaking with two Higgs doublets	1978	174.189
e	Theory of μ -Meson Decay with the Hypothesis of Nonconservation of Parity	1957	168.667
e	Neutrino masses in $SU(2) \otimes U(1)$ theories	1980	166.091
e	Gauge Dependence of the Wave-Function Renormalization Constant in Quantum Electrodynamics	1959	161.553
e	Binding Energy of the Positronium Molecule	1947	157.842
e	On Gauge Invariance and Vacuum Polarization	1951	155.702

e	Observables for the Analysis of Event Shapes in e^+e^- Annihilation and Other Processes	1978	155.518
e	On Bound States and Scattering in Positron Theory	1951	150.988
e	Structure of Phenomenological Lagrangians. II	1969	149.474
e	Quantum Electrodynamics on Null Planes and Applications to Lasers	1971	146.69
e	Bremsstrahlung, Synchrotron Radiation, and Compton Scattering of High-Energy Electrons Traversing Dilute Gases	1970	146.225
e	Asymptotic Behavior of Non-Abelian Gauge Theories to Two-Loop Order	1974	145.772
e	On a Theory of Particles with Half-Integral Spin	1941	140.75
e	On the Classical Radiation of Accelerated Electrons	1949	137.45
e	Multiquark baryons and the MIT bag model	1979	136.798
e	Erratum: Compton scattering. II. Differential cross sections and left-right asymmetry	1975	136.5
e	On the Possibility of Electric Dipole Moments for Elementary Particles and Nuclei	1950	135.899
e	Gravitational and Electromagnetic Mass in the Born-Infeld Electrodynamics	1935	135.632

コミュニティ	タイトル	出版年	<i>B</i>
f	Quantum statistical mechanics in a closed system	1991	439.226
f	Chaos and quantum thermalization	1994	334.727
f	Oxides Which Show a Metal-to-Insulator Transition at the Neel Temperature	1959	310.554
f	Interactions and the Anderson transition	1980	283.911
f	Information Theory and Statistical Mechanics	1957	256.688
f	Statistical Mechanics of the XY Model. I	1970	232.877
f	Statistical Mechanics of the XY Model. II. Spin-Correlation Functions	1971	214.853
f	Absence of Diffusion in Certain Random Lattices	1958	205.618
f	Metal-Insulator Transition	1968	165.125

コミュニティ	タイトル	出版年	<i>B</i>
g	Hard Superconductivity: Theory of the Motion of Abrikosov Flux Lines	1964	387.081
g	Magnetization of Hard Superconductors	1962	377.054
g	Magnetization of High-Field Superconductors	1964	328.585
g	Theory of Flux Creep in Hard Superconductors	1962	258.222
g	Surface Barrier in Type-II Superconductors	1964	236.894
g	Flux Creep in Type-II Superconductors	1969	214.115
g	Theory of the Motion of Vortices in Superconductors	1965	142.991
g	Inertial Mass of a Moving Fluxoid	1965	136.305

コミュニティ	タイトル	出版年	<i>B</i>
h	On Massive Neutron Cores	1939	587.444
h	Static Solutions of Einstein's Field Equations for Spheres of Fluid	1939	541.632
h	Can Gravitation Have a Finite Range?	1972	466.064
h	Relativistic Equations for Adiabatic, Spherically Symmetric Gravitational Collapse	1964	291.893
h	The Influence of the Expansion of Space on the Gravitation Fields Surrounding the Individual Stars	1945	288.625
h	On the Weight of Heat and Thermal Equilibrium in General Relativity	1930	261.305
h	The Particle Problem in the General Theory of Relativity	1935	229.759
h	Klein-Gordon Geon	1968	228.459
h	On the Problem of the Entropy of the Universe as a Whole	1931	226.0
h	Systems of Self-Gravitating Particles in General Relativity and the Concept of an Equation of State	1969	216.136
h	Temperature Equilibrium in a Static Gravitational Field	1930	205.268
h	An Undulatory Theory of the Mechanics of Atoms and Molecules	1926	202.806
h	Axisymmetric Black Hole Has Only Two Degrees of Freedom	1971	201.446
h	Gravitational Radiation from Point Masses in a Keplerian Orbit	1963	193.09
h	Infrared Photons and Gravitons	1965	184.373
h	Asymptotic Symmetries in Gravitational Theory	1962	181.549
h	Extraction of Energy and Charge from a Black Hole	1973	180.63
h	Gravitational Radiation and the Motion of Two Point Masses	1964	175.456
h	On Continued Gravitational Contraction	1939	173.709
h	Review of a Generalized Electrodynamics	1948	171.266
h	Low-Energy Theorem for Graviton Scattering	1968	166.783
h	Gravitational Collapse and Space-Time Singularities	1965	165.893
h	Canonical Variables for General Relativity	1960	163.55
h	Gravitational Field of a Spinning Mass as an Example of Algebraically Special Metrics	1963	160.694
h	Corrections and Additional Remarks to our Paper: The Influence of the Expansion of Space on the Gravitation Fields Surrounding the Individual Stars	1946	158.299
h	Klein-Gordon equation and rotating black holes	1980	157.016
h	Covariant Conservation Laws in General Relativity	1959	156.991
h	f-Dominance of Gravity	1971	156.441
h	Black hole in a uniform magnetic field	1974	152.433
h	A Generalized Electrodynamics Part II—Quantum	1944	151.208
h	Interaction of Neutrinos and Gravitational Fields	1957	148.157

h	Gravitational Radiation from Colliding Black Holes	1971	147.926
h	Renormalization of higher-derivative quantum gravity	1977	147.854
h	Dynamical Structure and Definition of Energy in General Relativity	1959	147.254
h	Equations of State of Elements Based on the Generalized Fermi-Thomas Theory	1949	143.794
h	Effective Potential for Even-Parity Regge-Wheeler Gravitational Perturbation Equations	1970	143.038
h	A Generalized Electrodynamics Part I—Non-Quantum	1942	141.277
h	Reality of the Schwarzschild Singularity	1968	140.302
h	Extension of the Low Soft-Photon Theorem	1968	137.239

コミュニティ	タイトル	出版年	<i>B</i>
i	Resonance Absorption by Nuclear Magnetic Moments in a Solid	1946	511.628
i	On the Process of Space Quantization	1936	442.346
i	A Molecular Beam Resonance Method with Separated Oscillating Fields	1950	365.367
i	Theory of Diffraction by Small Holes	1944	362.424
i	Space Quantization in a Gyration Magnetic Field	1937	355.034
i	Proceedings of the American Physical Society	1946	288.181
i	Phase Transition in the Dicke Model of Superradiance	1973	266.834
i	Coherence in Spontaneous Radiation Processes	1954	246.553
i	The Influence of Retardation on the London-van der Waals Forces	1948	242.82
i	Effects of Diffusion on Free Precession in Nuclear Magnetic Resonance Experiments	1954	231.147
i	Anomalous Diffraction Gratings	1935	218.967
i	Double Stern-Gerlach Experiment and Related Collision Phenomena	1932	218.452
i	Proceedings of the American Physical Society	1946	175.175
i	Visible Light from Localized Surface Charges Moving across a Grating	1953	175.082
i	Van der Waals forces and zero-point energy for dielectric and permeable materials	1974	170.343
i	Polarization of Nuclei in Metals	1953	163.93
i	Low field electron-nuclear spin coupling in gallium arsenide under optical pumping conditions	1977	159.314
i	On Nonadiabatic Processes in Inhomogeneous Fields	1937	159.171
i	Spin Echoes	1950	156.809
i	The Polarizabilities of Ions from Spectra	1933	152.874
i	Multiphoton Process Observed in the Interaction of Microwave Fields with the Tunneling between Superconductor Films	1963	152.528
i	Magnetic Resonance for Nonrotating Fields	1940	149.157
i	Steady States and Quasienergies of a Quantum-Mechanical System in an Oscillating Field	1973	148.222
i	Quantum Theory of Radiation	1932	146.373
i	Stark Effect in Rapidly Varying Fields	1955	136.541

コミュニティ	タイトル	出版年	B
j	Interaction between the d-Shells in the Transition Metals. II. Ferromagnetic Compounds of Manganese with Perovskite Structure	1951	1790.892
j	Considerations on Double Exchange	1955	1347.705
j	Neutron Diffraction Study of the Magnetic Properties of the Series of Perovskite-Type Compounds [(1 - x)La, xCa]MnO ₃	1955	1105.974
j	Effects of Double Exchange in Magnetic Crystals	1960	892.117
j	Theory of the Role of Covalence in the Perovskite-Type Man- ganites [La, M(II)]MnO ₃	1955	819.383
j	Finite Elastic Strain of Cubic Crystals	1947	428.107
j	Interaction Between the d Shells in the Transition Metals	1951	328.283
j	Antiferromagnetism. Theory of Superexchange Interaction	1950	270.868
j	Relationship Between Crystal Symmetry and Magnetic Prop- erties of Ionic Compounds Containing Mn ³⁺	1961	157.209
j	Criterion for Ferromagnetism from Observations of Magnetic Isotherms	1957	155.553
j	The Dirac Vector Model in Complex Spectra	1934	152.57
j	First-Order Localized-Electron ⇌ Collective-Electron Transi- tion in LaCoO ₃	1967	152.297
j	Coherent-Potential Approximation for a Nonoverlapping- Muffin-Tin-Potential Model of Random Substitutional Alloys	1972	145.431

コミュニティ	タイトル	出版年	<i>B</i>
k	Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?	1935	1392.824
k	The Uncertainty Principle	1929	753.772
k	Quantized Space-Time	1947	467.331
k	Proposed Experiment to Test Local Hidden-Variable Theories	1969	439.364
k	A Suggested Interpretation of the Quantum Theory in Terms of "Hidden" Variables. I	1952	305.016
k	Can Quantum-Mechanical Description of Physical Reality be Considered Complete?	1935	273.984
k	On the Problem of Hidden Variables in Quantum Mechanics	1966	241.888
k	A Suggested Interpretation of the Quantum Theory in Terms of "Hidden" Variables. II	1952	240.264
k	On the Quantum Correction For Thermodynamic Equilibrium	1932	230.243
k	On the Analogy Between Classical and Quantum Mechanics	1945	214.258
k	Hidden-Variable Example Based upon Data Rejection	1970	201.943
k	Hidden Variables, Joint Probability, and the Bell Inequalities	1982	181.41
k	Stochastic Dynamics of Quantum-Mechanical Systems	1961	158.826
k	Charge Superselection Rule	1967	152.705
k	Equivalence of Semiclassical and Quantum Mechanical Descriptions of Statistical Light Beams	1963	142.738
k	On Dissipative Systems and Related Variational Principles	1931	139.317

コミュニティ	タイトル	出版年	<i>B</i>
1	Instability of Electrified Liquid Surfaces	1917	295.462
1	Quantum Mechanics and Radioactive Disintegration	1929	273.116
1	The Calculation of Errors by the Method of Least Squares	1932	271.549
1	An Attempt to Calculate the Number of Energy Levels of a Heavy Nucleus	1936	239.745
1	Diatomic Molecules According to the Wave Mechanics. II. Vibrational Levels	1929	216.36
1	On the Displacement of Corresponding Energy Levels of C13 and N13	1951	213.648
1	Interpretation of Some of the Excited States of 4n Self-Conjugate Nuclei	1956	206.704
1	On the Quantum Theory of the Capture of Electrons	1928	191.528
1	The Threshold Law for Single Ionization of Atoms or Ions by Electrons	1953	180.345
1	Note on the Transmission and Reflection of Wave Packets by Potential Barriers	1932	172.88
1	On the Consequences of the Symmetry of the Nuclear Hamiltonian on the Spectroscopy of Nuclei	1937	169.787
1	On the Connection Formulas and the Solutions of the Wave Equation	1937	169.2
1	Lower Limit for the Energy Derivative of the Scattering Phase Shift	1955	161.393
1	On the Derivation of the Dispersion Formula for Nuclear Reactions	1939	157.155
1	On the Vibrations of Polyatomic Molecules	1932	156.767
1	The Mechanism of Nuclear Fission	1939	156.246
1	Interference in the Photo-Ionization of Molecules	1966	143.33
1	Lifetime Matrix in Collision Theory	1960	140.211
1	Synthesis of the Elements in Stars	1957	138.485
1	Diffraction in Time	1952	135.532