

博士論文

ダークウェブハイパーリンクグラフの
時系列視覚化によるダークウェブネット
トワークの変化の理解

Taichi AOKI

青木 太一

情報セキュリティ大学院大学

情報セキュリティ研究科

情報セキュリティ専攻

2022年3月

目次

第1章 序論	1
第2章 倫理	6
第3章 ダークウェブの探索	7
第1節 概要	7
第2節 関連研究	8
第1項 ワールドワイドウェブの構造からグラフを作り分析する研究	10
第2項 ダークウェブの構造からグラフを作り分析する研究	11
第3項 ダークウェブのフォーラム構造からグラフを作り分析する研究	13
第4項 ダークウェブの情報検索にグラフを用いて効率化する研究	14
第5項 ダークウェブ犯罪情報の視覚化を目的とする研究	15
第6項 ダークウェブ測定のガイドラインに関する研究	15
第7項 ダークウェブのネットワーク構造の時間的変化をグラフにより視覚化する観点に おける本研究の位置づけ	16
第3節 データセット構築の手法	17
第4節 データセット構築の結果	19
第1項 データセットの構築結果	19
第2項 スナップショットの生成結果	20
第3項 スナップショットの大きさ	21
第5節 データセットの評価	23
第1項 スナップショットの差分の視覚化	23
第2項 スナップショット間におけるドメインの揮発性	24
第3項 スナップショット間で変化するドメインの偏り	24
第4項 ドメインの生存日数	27
第6節 議論	31
第7節 本章のまとめ	35
第4章 ダークウェブネットワークの視覚化	37
第1節 概要	37
第2節 関連研究	37

第1項	ダークウェブの構造からグラフを作り分析する研究.....	37
第2項	ダークウェブのフォーラム構造からグラフを作り分析する研究	38
第3項	ダークウェブの情報検索にグラフを用いて効率化する研究.....	39
第4項	ダークウェブ犯罪情報の視覚化を目的とする研究.....	39
第5項	本研究の位置づけ.....	40
第3節	グラフによる視覚化の手法	41
第4節	グラフによる視覚化の結果	43
第1項	スナップショットAの視覚的標本の抽出.....	44
第2項	スナップショットBの視覚的標本の抽出.....	45
第3項	スナップショットCの視覚的標本の抽出.....	46
第4項	スナップショットDの視覚的標本の抽出.....	47
第5項	スナップショットEの視覚的標本の抽出.....	49
第5節	グラフによる視覚化の評価	51
第1項	視覚的に抽出した標本の特徴評価.....	51
第2項	グラフ中心性指標の導出とハブノードの抽出	63
第3項	グラフによる視覚化結果と中心性指標との比較.....	64
第6節	議論	68
第7節	本章のまとめ	70
第5章	ダークウェブネットワークのノード数の推定.....	71
第1節	概要	71
第2節	関連研究	71
第3節	手法	73
第4節	結果	73
第5節	評価	76
第6節	議論	80
第7節	本章のまとめ	80
第6章	結論.....	81

図表

図表 1 観測期間中のダークウェブの主なイベント	5
図表 2. 関連研究と本研究の位置づけ	9
図表 3. HTTP GET リクエストを使ったダークウェブからの HTML 情報の取得.	18
図表 4. Tor プロキシを経由して HTTP リクエストを発行するコマンド.	18
図表 5. HTTP レスポンスヘッダに含まれる Content-Type の例.	18
図表 6. HTML ファイルからリンクを抽出するコマンド.	18
図表 7. 観測期間中のクローリング数の 30 日幅のプロット	19
図表 8. 各スナップショットの収集期間と収集数	20
図表 9. スナップショット間の一意のドメイン数	21
図表 10. スナップショット間の一意のドメイン数	22
図表 11. スナップショット間の一意のドメイン数の変化	22
図表 12. スナップショット間のドメインの差分の視覚化	26
図表 13 データセットに含まれる全てのドメインの生存日数表	28
図表 14 データセットに含まれる全てのドメインの生存日数のプロット	29
図表 15 データセットに含まれる全てのドメインの生存率	29
図表 16 データセットから無作為にサンプリングした 100 ドメインの活性状態	30
図表 17. オニオンサービスの数	34
図表 18. スナップショット間で変化のあるドメインの割合	35
図表 19. ハイパーリンク関係の方向性の色による視覚化	42
図表 20. ダークウェブハイパーリンクネットワークグラフの視覚化	43
図表 21. スナップショット A のハイパーリンクネットワークグラフの視覚化	44
図表 22. スナップショット B のハイパーリンクネットワークグラフの視覚化	45
図表 23. スナップショット C のハイパーリンクネットワークグラフの視覚化	46
図表 24. スナップショット D のハイパーリンクネットワークグラフの視覚化	48
図表 25. スナップショット E のハイパーリンクネットワークグラフの視覚化	50
図表 26. 視覚的に抽出した 2 1 標本の RGB 成分構成比率	52
図表 27. 基準標本群の分布と分類	53
図表 28. 基準標本群の各標本に占める RGB 成分の比率	53
図表 29. スナップショット A から視覚的に抽出した各標本の RGB 成分	55
図表 30. スナップショット A から視覚的に抽出した各標本の分布	55

図表 31. スナップショット B から視覚的に抽出した各標本の RGB 成分	56
図表 32. スナップショット B から視覚的に抽出した各標本の分布	56
図表 33. スナップショット C から視覚的に抽出した各標本の RGB 成分	57
図表 34. スナップショット C から視覚的に抽出した各標本の分布	57
図表 35. ラベル c0 周辺の標本の分布	58
図表 36. スナップショット D から視覚的に抽出した各標本の RGB 成分	59
図表 37. スナップショット D から視覚的に抽出した各標本の分布	59
図表 38. ラベル d4・d5 周辺の標本の分布	60
図表 39. スナップショット E から視覚的に抽出した各標本の RGB 成分	61
図表 40. スナップショット E から視覚的に抽出した各標本の分布	61
図表 41. ラベル e1・e3・e4 周辺の標本の分布	62
図表 42. 次数中心性の上位 10 ドメインの占めるカテゴリの集計	65
図表 43. 出力次数中心性の上位 10 ドメインの占めるカテゴリの集計	65
図表 44. 入力次数中心性の上位 10 ドメインの占めるカテゴリの集計	66
図表 45. ページランクの上位 10 ドメインの占めるカテゴリの集計	66
図表 46. 標識再捕獲法により推定されたダークウェブのドメイン数	74
図表 47. 標識再捕獲法における再捕獲率	75
図表 48. 本実験におけるダークウェブの大きさの推定値と誤差の範囲	78
図表 49. 本実験におけるダークウェブの大きさの推定に使用した値	79
図表 50. ばねモデルによるグラフィレイアウトのイメージ	91
図表 51. サークルモデルによるグラフィレイアウトのイメージ	92
図表 52. 階層モデルによるグラフィレイアウトのイメージ	93
図表 53. 階層モデルによるグラフィレイアウトの拡大図	93
図表 54. スナップショット A の入力次数の上位 10 ノード (n=48360)	94
図表 55. スナップショット B の入力次数の上位 10 ノード (n=53300)	94
図表 56. スナップショット C の入力次数の上位 10 ノード (n=70684)	95
図表 57. スナップショット D の入力次数の上位 10 ノード (n=103674)	95
図表 58. スナップショット E の入力次数の上位 10 ノード (n=54509)	95
図表 59. スナップショット F の入力次数の上位 10 ノード (n=17464)	96
図表 60. スナップショット A-F の入力次数の上位 10 ノード (n=172740)	96
図表 61. スナップショット A の出力次数の上位 10 ノード (n=48360)	97
図表 62. スナップショット B の出力次数の上位 10 ノード (n=53300)	97
図表 63. スナップショット C の出力次数の上位 10 ノード (n=70684)	98
図表 64. スナップショット D の出力次数の上位 10 ノード (n=103674)	98
図表 65. スナップショット E の出力次数の上位 10 ノード (n=54509)	98
図表 66. スナップショット F の出力次数の上位 10 ノード (n=17464)	99

図表

図表 67. スナップショット A-F の出力次数の上位 10 ノード (n=172740)	99
図表 68. スナップショット A の PageRank の上位 10 ノード (n=48360)	100
図表 69. スナップショット B の PageRank の上位 10 ノード (n=53300)	100
図表 70. スナップショット C の PageRank の上位 10 ノード (n=70684)	101
図表 71. スナップショット D の PageRank の上位 10 ノード (n=103674)	101
図表 72. スナップショット E の PageRank の上位 10 ノード (n=54509)	101
図表 73. スナップショット F の PageRank の上位 10 ノード (n=17464)	102
図表 74. スナップショット A-F の PageRank の上位 10 ノード (n=172740)	102
図表 75. 標本 A の色成分.....	103
図表 76. 標本 B の色成分.....	104
図表 77. 標本 C の色成分	104
図表 78. 標本 D の色成分.....	105
図表 79. 標本 E の色成分.....	105
図表 80. 標本 F の色成分.....	106
図表 81. 標本 G の色成分.....	106
図表 82. 標本 H の色成分.....	107
図表 83. 標本 I の色成分.....	107
図表 84. 標本 J の色成分.....	108
図表 85. 標本 K の色成分	108
図表 86. 標本 L の色成分.....	109
図表 87. 標本 M の色成分.....	109
図表 88. 標本 N の色成分	110
図表 89. 標本 O の色成分.....	110
図表 90. 標本 P の色成分.....	111
図表 91. 標本 Q の色成分.....	111
図表 92. 標本 R の色成分.....	112
図表 93. 標本 S の色成分.....	112
図表 94. 標本 T の色成分	113
図表 95. 標本 U の色成分	113
図表 96. 標本 a0 の色成分	114
図表 97. 標本 a1 の色成分	114
図表 98. 標本 a2 の色成分	115
図表 99. 標本 a3 の色成分	115
図表 100. 標本 a4 の色成分	116
図表 101. 標本 a5 の色成分	116
図表 102. 標本 a6 の色成分	117

図表 103. 標本 a7 の色成分	117
図表 104. 標本 b0 の色成分	118
図表 105. 標本 b1 の色成分	118
図表 106. 標本 b2 の色成分	119
図表 107. 標本 b3 の色成分	119
図表 108. 標本 b4 の色成分	120
図表 109. 標本 c0 の色成分	121
図表 110. 標本 c1 の色成分	121
図表 111. 標本 c2 の色成分	122
図表 112. 標本 c3 の色成分	122
図表 113. 標本 c4 の色成分	123
図表 114. 標本 c5 の色成分	123
図表 115. 標本 d0 の色成分	124
図表 116. 標本 d1 の色成分	124
図表 117. 標本 d2 の色成分	125
図表 118. 標本 d3 の色成分	125
図表 119. 標本 d4 の色成分	126
図表 120. 標本 d5 の色成分	126
図表 121. 標本 e0 の色成分	127
図表 122. 標本 e1 の色成分	127
図表 123. 標本 e2 の色成分	128
図表 124. 標本 e3 の色成分	128
図表 125. 標本 e4 の色成分	129
図表 126. ダークウェブハイパーリンクネットワークグラフの拡大図	130

用語

本稿で使用する主要な用語を説明する。

【匿名通信システム】

匿名性を保つことを目的とする匿名通信を実現するシステム。このような匿名通信システムには、Tor (The Onion Router) や I2P (The Invisible Internet Project) などが存在する。ダークウェブとしては Tor がよく使われているため、本研究では主に Tor ネットワークを対象とする。

【Tor】

Tor (The Onion Route) は、米海軍調査研究所の支援によって開発されたオニオンルーティングによる匿名化技術のリファレンス実装である。現在では電子フロンティア財団によって支援され、一般に使われるようになった 2004 年から現在まで広く受け入れられていることから、堅牢な匿名システムといえよう。発信者の匿名化には、古くは Proxy 技術が使われており、Tor はその Proxy 技術を大きく発展させた多段 Proxy システムの延長線上にある技術と捉えて良い [1]。Tor の中継ノードは SOCKS Proxy として動作することにより TCP 通信を中継することが可能である。オニオンルーティングの名称の元となる玉ねぎの皮のように、中継ノード (Onion Router: OR) を経由するたびに通信内容を多重に暗号化及び復号しながら転送する仕組みが Tor の大きな特徴である。暗号化されたパケットを中継している OR ノードは、自身の前後の通信相手しか知ることが出来ないことに加え、中継ノードに対する本来の発信者と受信者の情報は末端同士で共有する鍵によって暗号化することで通信経路上の匿名性を維持している。Tor は TCP ベースの匿名サービスを提供する機能があり、Hidden Service と呼ばれる。Hidden Service では固定長の半角英数字で構成される文字列に末尾に .onion を付けたドメインがサービスを一意に識別するアドレスとして用いられる。Hidden Service ノードの名前解決には P2P の技術 (分散ハッシュテーブル: DHT) をベースとしたアドレス解決機能が使われている。

【I2P】

Tor の発展形として I2P (The Invisible Internet Project) が提案されており、ルータとしての実装が提供されている。I2P では暗号化の仕組みを Garlic encryption (ガーリック暗号化) と呼び、Tor で使用されるタマネギの例えを意識したネーミングがされている。匿

名ネットワークとしての多くの仕組みは Tor で用いられる方法を踏襲しながらも、Tor との大きな差は、通信に使用されるトンネルを 1 方向に限定し、送信と受信に関する経路を分離したことに加え、パケット長を固定長とした上で、1つの暗号化メッセージ内に複数の送信メッセージが含まれることで、相関分析等による識別攻撃からの耐性を高める工夫がなされている。通信に使用されるトンネルは2種類あり、クライアントトンネル（メッセージ通信）と、探索用トンネル（探索・管理用通信）が用意されている。現時点において I2P の実装は技術的な点から利用上の障壁が高く、サイバー犯罪への利用は限定的である。

【ダークウェブ】

ダークウェブは、犯罪などの目的のために使用されるサイバー空間上の領域で、登場初期から現在においてワールドワイドウェブと同じ技術が使われている。インターネットを使用したオーバーレイネットワークとして作られることが多く、アクセスするために特定のソフトウェアやシステム構成、または承認が必要な領域に存在している。ダークウェブを介することによって、利用者に関する IP アドレス情報を秘匿した匿名で相手との間の通信を行い、違法なビジネスが行われる。ダークウェブの言葉の登場に際してその概念を表現した説明 [2]により、ダークウェブは検索エンジンによってインデックス化されず、検索できない領域であると信じられている。

【グラフ】

ネットワークのつながり方に着目して抽象化された「点とそれらを結ぶ線」の概念をグラフと呼び、主にネットワーク科学の分野で使われる。グラフは物理学的、生物学的、社会的、及び情報システムにおける多くの種類の関係と過程をモデル化するために使うことができ、多くの現実的問題はグラフによって表わすことができる。現実世界のシステムへの応用を強調する時には、「ネットワーク」という用語がグラフを意味するために使われ、現実世界のシステムをネットワークとして表現し理解する研究分野はネットワーク科学と呼ばれる。ネットワーク科学では、ネットワーク全体ないしは局所を具体的に解析する方法が道具立てとして用いられ、ネットワークの性質を表す指標が提案されている。

【中心性指標】

ネットワークの性質を表す指標の一つに中心性指標がある。ネットワークにおいてどの頂点が要であるかを知りたい場合、要である度合いを頂点の中心性と呼ばれる。古くは社会学ネットワーク分析の分野で用いられ、代表的な指標は 1980 年頃には既に出揃ったとされる。現在では複雑ネットワークの分野においても同様の指標が用いられている。本研究では中心性指標として代表的な次数中心性と、ワールドワイドウェブの検索エンジンにおけるランク付けに使われたことで脚光を浴びた PageRank を使用して、視覚化による傾向と比較する。

【標識再捕獲法】

標識再捕獲法 (mark-recapture) は個体群生態学の分野で使用され、個体群を構成する個体数を推定する為の方法である。個体群生態学においては、あるフィールドに棲息する野生動植物の全個体を把握してその個体数の全てを数えることは不可能に近い為、推定によって個体数を予想する必要に迫られる。本研究では、標識再捕獲法として代表的な Petersen によるモデルを使用してダークウェブの大きさを推定する。

用語の定義

本稿では特に断りがない限り、以下の用語については下記のように定義し使用する。

【匿名通信システム】

匿名性を保つことを目的とする匿名通信を実現するシステムを指す。このような匿名通信システムには、Tor (The Onion Router) や I2P (The Invisible Internet Project) などが存在する。ダークウェブとしては Tor がよく使われている。

【匿名ネットワーク】

匿名性を保つことを目的とする匿名通信システムにより作られるネットワークを指す。具体的には、Tor や I2P により作られるネットワークのこと。Tor や I2P はインターネットを介して接続し合い、送受信者の匿名性を保ちながら TCP パケットを転送することが可能となるように設計されている。TCP のみにより動作するアプリケーションは、Tor ネットワークで利用することが出来る。

【Tor ネットワーク】

匿名通信システムの1つである Tor により作られる匿名ネットワークのことを指す。

【ワールドワイドウェブ】

インターネットを通じて公開されたウェブページが相互に接続されたシステムを指す。

【ウェブ】

ワールドワイドウェブと同義である。

【ダークウェブ】

Tor ネットワークで http サーバーが稼働して html によりウェブとして繋がり合う領域のことを指す。ウェブと同様に、URL とドメインによりリソースが特定される。

【ダークウェブネットワーク】

ダークウェブと同義であるが、特にネットワークに注目する場合に使用する。

【ダークウェブ http サーバー】

ダークウェブと同義であるが、特に http サーバーに注目する場合に使用する。

【グラフ】

単独でグラフと呼ぶ場合、グラフ理論におけるグラフ構造を指す。

【グラフデータ】

グラフ構造を構成するための情報を指す。

【グラフィアウト】

グラフ構造を視覚化するための画面整列手法を指す。

【ノード】

ネットワークを構成するシステムの個々を指す。

【エッジ】

グラフ構造においてノード同士の接続関係を示す辺を指す。

【ホスト】

Tor ネットワークで使用される URL におけるホスト名までを含めたドメインに相当する部分で識別されるノードを指す。

【ドメイン】

Tor ネットワークで使用される URL の内、トップレベルドメインとセカンドレベルドメインの組み合わせにより識別されるノードを意味する。Tor ネットワークにおけるトップレベルドメインは.onion が使用される。Tor ネットワークにおけるセカンドレベルドメインは、Tor ネットワークにおいてノードを一意に識別するために使用され、Tor ネットワークへの登録が要求される。

【ダークウェブの大きさ】

ダークウェブネットワークを構成する一意のドメインの数と同義として使用する。

【フォーラム】

人間同士が情報交換を行う目的で設置されるウェブサービスを指す。

第1章

序論

武器や麻薬の密売などのさまざまな違法行為に関するコンテンツがダークウェブで配布されている。これらの違法な活動は、Tor [3]やI2P [4]の匿名ネットワークによって利用者の匿名性が保たれている。通常、これらの匿名ネットワークにはワールドワイドウェブから直接アクセスすることはできず、専用のソフトウェアを使用するとアクセスできる。

ダークウェブを理解することにより、サイバーセキュリティにおけるさまざまな戦略の検討に役立てることが出来る。たとえば、ダークウェブで取り交わされる脅威の傾向を知ることが出来る。企業や政府がサイバーセキュリティ上の防衛手段の適切さを検討するための指標が得られる。他にも、ダークウェブネットワークの状態を知ることから、ダークウェブの活性状態を予測できる可能性もある。そして、ダークウェブの活性状態からは、観測されたダークウェブ以外の未知のネットワークに対して、違法活動をするアクターの流入や流出を予測できる可能性もある。このように、ダークウェブに集中している潜在的な脅威を理解することは、サイバーセキュリティ上の防衛戦略にとって役に立つ。

本研究では、観測時点においてダークウェブとして利用者の多い Tor を対象とした分析を行う。本研究によるダークウェブの観測期間中に、ダークウェブではいくつかの重要なイベントが発生した [図表 1]。2019年4月30日、当時最大のダークマーケットであったドリームマーケット(Dream Market)が閉鎖された [5]。2019年4月6日、ダークウェブ上の大規模なフォーラムであるギャングスターズパラダイスフォーラム(Gangsta's Paradise Forum)が閉鎖された [6]。2019年5月2日、ドイツの警察はダークウェブで当時二番目に大きなマーケットであるウォールストリートマーケット(Wall Street Market)をテイクダウンした [7]。2019年5月3日、ユーロポール(EUROPOL)はダークウェブで当時三番目に大きなマーケットであったヴァルハラマーケット(Valhalla Marketplace)を閉鎖した [8]。2019年5月7日、国際的な協力により法執行機関がダークウェブニュースサイトの DeepDotWeb を閉鎖した [9]。2019年5月9日、ダークウェブニュースサイトである DarkWebNews が法執行機関による摘発を恐れて自らの手で閉鎖された [10]。2019年5月22日、オランダの警察はビットコインミキシングサイトである Bestmixer.io を閉鎖し

た [11]. はたしてこれらのイベントはダークウェブに対して有効な打撃を与えただろうか? これらの法執行機関による努力にも関わらず, ダークウェブは依然として存在し続け, 犯罪者はますますダークウェブを利用し, 違法なビジネスはその規模を拡大し続けている.

本研究は, ダークウェブネットワークの変化を視覚化することによって, ダークウェブネットワークの変化と変遷についての理解を深める手段を提供することを目的としている. 我々の知る限り, ダークウェブネットワークは巨大で複雑で流動的であるため, ダークウェブの状態を理解することは簡単ではない. ダークウェブの全てを瞬時にキャプチャすることは困難であり, ダークウェブの全てを一人の人間が追いかけて理解することは困難である. 本研究で披露するダークウェブネットワークの視覚化は, 大規模で複雑なネットワークがどのように変化してきたかという問題に対して直感的な理解を提供し, そこから現在と未来に向けた様々な洞察を得るための手段として役に立つ. 本研究の貢献は, ダークウェブから大局的な視点で変化についての洞察を得るための視覚化による手段を提供したことである.

本研究では, ダークウェブからの観測により収集した html テキストデータからハイパーリンクを抽出し, ダークウェブにおけるハイパーリンク関係のグラフデータを構築した. この観測データは, 本稿記載時点でダークウェブから抽出されたものの中では, 長期間安定的に記録され続けた最大規模のデータセットである. このデータセットから作られるグラフデータは時系列ごとに分割され視覚化を行った. 本研究は, ダークウェブのグラフデータを時系列の視点から視覚的に分析を行う最初の研究である. 視覚化されたグラフデータは, 視覚化によるメリットを明らかにするために, グラフから計算により自動的に求めることができる代表的な中心性指標と比較を行った. 本研究は, 視覚化されたグラフデータと中心性指標とを比較した最初の研究である. 更に, データセットに含まれるドメインテキストに対して, スナップショット間の差分を取ることで視覚化を行い, ダークウェブのノードが一部のフィッシングアクターにより大規模に占拠されていない, つまり匿名性という観点からは望ましい状態であることを確認した. Tor ネットワークにおいて, ドメインは公開鍵暗号により生成されるため, 恣意的に鍵を生成しない限り暗号学的な精度のランダム性をもってドメインが生成されることから, 変化の数が多くなるにつれ, スナップショット間のドメインテキストの差分の視覚化による変化箇所は均一に分布するよう見える. この状態は匿名性の点からは望ましい. 一方で, ダークウェブのドメインは, 総当たりによって意図的に取得することができるため, Tor ネットワークはフィッシングに対して脆弱である. この状態ではダークウェブの多くの部分は人間を騙す目的で類似するドメインが並ぶため, そのようなドメイン群が短期間に登場すると, 均一に塗りつぶされた結果が表示される. このような状態は匿名性の点からは脆弱である. 本研究は, ダークウェブのドメインの差分を時系列で視覚化した最初の研究である. 本研究では, データセットに含まれる観測されたドメインに対して生存期間を調査した. 本研究は, ダークウェブの大規模なドメインの生存期間についての統計情報を示す最初の研究である. 最後に, データセ

ットに含まれる観測データから、ダークウェブの大きさについて標識再捕獲法を用いて推定を試みた。本研究は、ダークウェブの大きさについて標識再捕獲法を用いて推定する最初の研究である。本研究では、以降特に断りがない限り、ダークウェブについてはTorネットワークでhttpサーバーが稼働してhtmlによりウェブとして繋がり合う領域のことを指す。

第2章では、倫理について記載する。ダークウェブへのアクセスとそこから引き起こされる一連の行動は、コンピュータ技術上の危険だけではなく、犯罪へ巻き込まれる・犯罪を誘発する・社会を混乱させる等の危険を伴うため、注意深く行う必要がある。本研究は、サイバー空間及び国や社会の安全と安定の確保を目的として行われ、法執行機関の協力を得て、技術的な安全性以外にも、適法性及び倫理について細心の注意を払った上で実施した。本研究により取り扱うデータセットは、ダークウェブの詳細な理解と、ダークウェブのナビゲーションが可能なコンテンツを有しているが、その扱いには細心の注意を払っている。前述の通り、これらの能力はダークウェブの研究にとっての基礎データとなりうるが、悪意を持ってダークウェブを利用したい者にとっても有益な情報であるため、違法行為に対する幫助となりうる可能性の懸念からも、安易に公開されるべき性質のものではない。本研究で取り扱う情報は、我が国のサイバーセキュリティ研究の発展と社会安定の目的のため、一般財団法人日本サイバー犯罪対策センターと法執行機関の協力を得て適法性が保たれるよう扱われ、その取り扱いを制限している。また、本稿執筆においても、ダークウェブの詳細なディテールの記述については必要最小限となるよう細心の注意を払っている。

第3章では、ダークウェブの探索について記載する。本研究では、観測時点においてダークウェブとして人気のあるTorネットワークを対象に観測を行った。本稿では、特に注釈が無い限り、ダークウェブとはTorネットワークであるとして扱う。ダークウェブは、その言葉の登場にあたり、検索エンジンによる検索では発見できないウェブの領域として位置づけられているため、自身により観測手段を確保する必要がある。本研究では、ダークウェブはhtmlにより繋がるネットワークである点に着目して、ウェブの真の目的は人間に対する情報とコミュニティへの誘導の提供であるとの仮説の元、ワールドワイドウェブと同様の手法でダークウェブの観測を行った。つまり、人間がダークウェブにアクセスしてコミュニティを追いかけ続けることが出来るのであれば、同様の方法で機械的に追い続けることが出来るというウェブの設計思想がダークウェブにおいても同様に働くであろうとの仮説に基づく。この仮説を証明することは困難であるが、本研究の観測においてはこの仮説は有効に機能し続けた。つまり、クローリングによるダークウェブ探索の動作が途中で途切れ、観測が停止することはなかった。本研究における観測により、おおよそ2年に渡る2018年6月1日から2021年1月30日までの期間において、25,270,157ページのhtmlテキストファイルを収集した。これらのテキストファイルは、幅優先探索によりTorネットワークから収集された。これらのテキストファイルからハイパーリンクを抽出し、グラフデータを構築した。そして、データセットは180日の期間により時系列ごとに分割

し、分析のためのダークウェブのスナップショットとした。これらのスナップショット間において、ドメインテキストの差分の視覚化を行い、ダークウェブのドメイン名の分布や生成・消失について偏りがみられないかを調べた。更に、ダークウェブのドメインの生存期間を調べ、統計情報を提供する。本章の結果として、ダークウェブのネットワークはドメイン名の観点から偏りは確認できなかった。そして、ダークウェブにおけるドメインの生存期間から求まるダークウェブの半減期についてもワールドワイドウェブと比べて1.5%~2.3%しかなく、著しく短いことが示された。

第4章では、ダークウェブネットワークの視覚化について記載する。第3章で作成したスナップショットの有向グラフデータについてばねモデルを用いて視覚化する。視覚化において、エッジの始点と終点でそれぞれ異なる色をつけることで、本研究特有の視覚化効果を得た。この視覚効果からは、ダークウェブグラフにおけるハブノードの存在と、低い次数で多くのノードが結合する領域の存在を浮き彫りにする。これらの視覚化結果を、ネットワークグラフの代表的な中心性指標と比較を行った。中心性指標とは、ネットワークにおける重要なノードを計算により識別するために使用される値であり、グラフデータから計算により自動的に求めることが出来る。結果として、中心性指標からは特徴があるとみなされないいくつかの領域を視覚化により発見することができた。更に、これらの視覚化結果を時系列で比較した結果、ダークウェブにおける情報検索の在り方が、リンクコレクションを中心とした関係性をもつ構造から、検索エンジンと相互参照を中心とする関係性をもつ構造へと徐々に変化している様子を示した。視覚化により明らかとなった領域からは、ダークウェブの中でも大きな勢力を持つボルノネットワークやフィッシングネットワークの存在が確認できた。分析の結果、これらの勢力は観測を開始した当初は存在せず、2019年5月以降に誕生したことが明らかとなった。

第5章では、ダークウェブネットワークのノード数の推定について記載する。ダークウェブの大きさを求めることは困難であると考えられており、本研究においてもダークウェブの全てを瞬時に観測することは実現できていない。そこで、本研究では観測データをもとに、Peterson による標識再捕獲法を用いてダークウェブの大きさの推定を試みた。Peterson による標識再捕獲法は個体群生態学の分野で野生動植物の個体数の推定を行う場合にしばしば使用される手法である。野生動植物の個体数も、あるフィールドにおいて全ての個体を同時に識別してカウントすることは困難であるため、ダークウェブの大きさの推定と同質の問題を抱えている。本研究の推定では、観測期間中のさまざまな時点で32回の測定を行い、ダークウェブネットワークの大きさは14,509から96,034であるとの推定値が示された。データセット全体の推定値平均は40,848であった。

第6章では、本研究における視覚化によって明らかとなったダークウェブの変化について結論付ける。本研究が視覚的にダークウェブの変化を示すまで、ダークウェブがどのように変化してきたかを知る手段は無かった。本研究で示した視覚化は、ダークウェブという巨大なネットワークがどのように動き、変化してきたかを描写し、過去と現在のダークウ

ウェブの姿、そして将来のダークウェブの変化についての洞察を得るための手段を提供する。本研究の観測期間において起こったダークウェブの最大の変化は、ダークウェブの中の情報検索のあり方が、リンク集から検索エンジンへ変化しつつあることを示唆する結果を得たことである。また、視覚的な情報からは、ネットワーク中心性指標では見えづらい領域の存在を発見できることもわかった。これらの領域は少ないリンクによってお互いが弱く結合するため、次数をもとにランクを求めるネットワーク中心性指標では数値が低くなっていた。

最後に、付録としていくつかのグラフィアウトの描画イメージと、本稿で使用したスナップショットにおけるネットワーク中心性指標の上位10ノードの値、目視により抽出した標本のRGB各成分の分解と、比較のために Watershed 法により自動認識された領域を掲載する。

図表 1 観測期間中のダークウェブの主なイベント

時期	イベント
2019年4月30日	当時最大のダークマーケットであったドリームマーケット (Dream Market)が閉鎖 [3].
2019年4月6日	ダークウェブ上の大規模なフォーラムであるギャングスターズパラダイスフォーラム(Gangsta's Paradise Forum)が閉鎖 [4].
2019年5月2日	ドイツの警察はダークウェブで当時二番目に大きなマーケットであるウォールストリートマーケット(Wall Street Market)をテイクダウン [5].
2019年5月3日	ユーロポール(EUROPOL)はダークウェブで当時三番目に大きなマーケットであったヴァルハラマーケット(Valhalla Marketplace)を閉鎖 [6].
2019年5月7日	国際的な協力により法執行機関がダークウェブニュースサイトの DeepDotWeb を閉鎖 [7].
2019年5月9日	ダークウェブニュースサイトである DarkWebNews が法執行機関による摘発を恐れて自らの手で閉鎖 [8].
2019年5月22日	オランダの警察はビットコインミキシングサイトである Bestmixer.io を閉鎖 [9].

第2章

倫理

ダークウェブに技術的にアクセス可能であるからといって、安易に近づくことは大きな懸念がある。それは決して興味本位であってはならない。ダークウェブには、コンピュータウイルスやコンピュータに危害を加える攻撃コードのような技術上の問題だけではなく、銃器・麻薬・児童ポルノのように所持することにより違法性が問われるコンテンツが存在し、そのコンテンツの周辺には、違法なビジネスを行うコミュニティと経済がある。ダークウェブへのアクセスとそこから引き起こされる一連の行動は、技術上の危険だけではなく、犯罪へ巻き込まれる・犯罪を誘発する・社会を混乱させる等の危険を伴うため、注意深く行う必要がある。

本研究は、我が国におけるサイバー空間及び国や社会の安全と安定の確保を目的として行われ、一般財団法人日本サイバー犯罪対策センター¹ や法執行機関の協力のもとで、技術的な安全性以外にも、適法性及び倫理についてそれぞれ有識者に助言を仰ぎ、注意深く実施した。

本研究により取り扱うデータセットは、ダークウェブの詳細な理解と、ダークウェブのナビゲーションが可能なコンテンツを有しているが、その扱いには細心の注意を払っている。前述の通り、これらの能力はダークウェブの研究にとっての基礎データとなりうるが、悪意を持ってダークウェブを利用したい者にとっても有益な情報であるため、違法行為に対する幫助となりうる可能性の懸念からも、安易に公開されるべき性質のものではない。本研究で取り扱う情報は、我が国のサイバーセキュリティ研究の発展と社会安定の目的のため、一般財団法人日本サイバー犯罪対策センターと法執行機関の協力を得て適切に管理され、その取り扱いを制限している。また、本稿執筆においても、ダークウェブの詳細なディテールの記述については必要最小限となるよう細心の注意を払った。

¹ 一般財団法人日本サイバー犯罪対策センター (<https://www.jc3.or.jp/>)

第3章

ダークウェブの探索

第1節 概要

ダークウェブの視覚化にあたり、元となるデータの収集を行った。ダークウェブは検索エンジンによる検索では発見できないウェブの領域として位置づけられている [2]ため、本研究においては Tor ネットワークに対してウェブと同様にクローリングによるダークウェブネットワークノードの全探索を行った。本研究で収集したデータセットは、ワールドワイドウェブで行われるものと同様の一般的なクローリング手法を用いて、Tor ネットワークから収集したデータを分析したものである。本研究のクローリングは 2018 年 6 月 1 日から 2021 年 1 月 30 日まで継続的に行われた。既存研究のダークウェブのクローリングでは、数日から数か月程度の短期間の収集が主であり、既存の Tor データセットとしては DUTA-10K [12]があるが、このデータセットは収集ボリュームに焦点が当てられており、収集された時期については判然としないため、収集時期に焦点を当てる研究には不向きである。また、DUTA-10K データセットに含まれる一意のドメイン数は 10,367 であり、本研究で収集した一意のドメイン数 172,740 と比べても極めて小さい。さらに、本研究で収集した期間を含む長期間継続的に収集されたダークウェブのデータセットは既存研究には存在しない。そのため、本章では、視覚化に先立ち時系列分析が可能なデータセットを構築する。

第2節 関連研究

本研究では、ダークウェブからクローリングしたデータを使い、時系列ハイパーリンクグラフ構造を作り、視覚化を行う。そのため、ダークウェブのネットワーク構造についてグラフ分析と視覚化を行う研究を中心に関連研究の調査を行った。図表 2 に既存の調査を要約し、本研究を位置付けるとともに、比較できるようにした。

ウェブが登場してから 2000 年頃の間、ワールドワイドウェブのネットワーク構造についていくつかの研究が行われてきた [13] [14] [15] [16]。なかには、ネットワークグラフを介したワールドワイドウェブの視覚化を取り扱う研究もみられる [17]。ワールドワイドウェブと同様に、ダークウェブも http によりネットワークが作られているため、ウェブに適用される手法のいくつかは、ダークウェブでも使用できる。ダークウェブもウェブと同様に、登場以来広く研究対象となっている。ダークウェブにおいて人気のある研究領域は、サイバー犯罪に関するソーシャルネットワークとフォーラムネットワークの分析 [18] [19] [20] [21]、そしてダークウェブネットワークのグラフ構造の分析 [22] [23] [24] [25] [26] [27] [28] である。これらの研究の中においてダークウェブの広範なネットワーク構造の視覚化に焦点を当てた研究は Brunner ら [22] によるもの、Domenico ら [24] によるもの、Cilleruelo ら [28] によるものの 4 つがある。

図表 2. 関連研究と本研究の位置づけ

関連研究	調査対象	データセット サイズ	関連性	視覚化対象	時系列 視覚化	ノード 数推定
Broder ら [15]	WWW	1,000,000,000	Hyperlink	Conceptual diagram	No	No
Toyoda ら [14]	WWW	2,830,000,000	Hyperlink	Graph	Yes	No
Heer ら [27]	WWW	n/a	Social network	Graph	No	No
Fu ら [19]	Tor	650	Social network	Graph	No	No
Hyperion ら [29]	Tor	3,700	Contents structure	Graph	No	No
Domenico ら [24]	Tor	5,535	Hyperlink	Graph	No	No
Griffith ら [23]	Tor	13,117	Hyperlink	Chart	No	No
Brunner ら [22]	Tor	34,714	Hyperlink	Graph	No	No
Cilleruelo ら [28]	Tor and I2P	46,562	Hyperlink	Graph	No	No
Our study [30] [31]	Tor	172,740	Hyperlink	Graph	Yes	Yes
Park ら [25]	Dark web	n/a	n/a	Chart	No	No
Zhou ら [18]	Dark web	n/a	Social network	Graph	No	No
Yang ら [20] [21]	Dark web	n/a	Social network	Graph	No	No
Sallaberry ら [13]	Information search	n/a	Search query	Graph	No	No

第1項 ワールドワイドウェブの構造からグラフを作り分析 する研究

Broder ら [15]は、ワールドワイドウェブから収集された 10 億を超えるノードからグラフを作成し、その構造を調査した。彼らはウェブのグラフ分析のための方法を開発した。彼らはその研究の中で、グラフとしてのウェブの研究は、それ自体が魅力的であるだけでなく、クローリング、検索、コミュニティの発見のためのウェブアルゴリズムとその進化を特徴づける社会学的現象に関する貴重な洞察を生み出すと述べている。彼らは 1999 年 5 月から 1999 年 10 月にかけてクローリングにより 200 万ページ以上のリンクを辿り、抽出された 15 億のリンクを使用してウェブのグラフを作った。彼らの提供する視覚化は、おそらく当時のコンピュータの性能がこの問題に対する視覚化のためには著しく不足していたために、概念図に留まっていた。彼らは測定したウェブ上のノード数を報告したが、ウェブのノード数の推定は行わなかった。

Toyoda ら [14]は、時系列グラフを使用してワールドワイドウェブのリンク構造を視覚化する方法を開発した。彼らは 1999 年から 2003 年にかけて日本のウェブを対象として 6 回のクローリングを行い、それらの時系列グラフを分析し、視覚化した。彼らはウェブの構造変化を追跡することについて、ウェブページの出現や消失とその関係性の変化を明らかにすることは、検索エンジンによるランク付けを適切にするためのリンクスパムの調査や、社会学的な研究としての文化の追跡に重要であると述べている。彼らの視覚化では、異なる時間のウェブの状態を 1 つの画面で同時に見る事が出来る工夫をしている。更に、ウェブページの出現や消失ごとに色分けがなされ、ウェブページがどの時点で変化したかを分かり易く視覚化した。彼らの研究では、ウェブの個々の詳細な変化について焦点を当てて分析がなされているものの、ウェブ全体について大局的な変化については議論されなかった。

Heer ら [27]は、ソーシャルネットワークグラフの構造を視覚化した。さらに、彼らは写真とグラフを組み合わせる方法を考案した。彼らはソーシャルネットワーク上の人物の探索とナビゲーションのために 2003 年にソーシャルネットワークサービス上からクローリングにより収集された 150 万ユーザーから構成されるグラフ構造を対象に、視覚化する方法を提供した。彼らの視覚化は写真とグラフにより表現され、複数のコミュニティに共通する人物を視覚的に見つけ出せる工夫がなされていた。彼らの視覚化には、ばねモデルが使われている。彼らはばねモデルを採用する理由として、接続の強さに応じてコミュニティに自動的に分類できるメリットがあると述べている。彼らの視覚化はインタラクションが重視されていたため、小規模な範囲の可視化に留まっていた。つまり、ソーシャルネットワーク全体を可視化してはいなかった。

ここで挙げた3つの研究は、いずれもダークウェブについては視覚化していない。そのため、ダークウェブについても同様の観点から研究を進めることは有益である。

第2項 ダークウェブの構造からグラフを作り分析する研究

Domenico ら [24]は、2013年12月、2014年5月、2015年1月の3つのデータセットを使用してダークウェブのネットワーク構造を調査し、それぞれ5,921、4,953、5,535ノードを分析した。彼らの研究は、複数のデータセットの比較を試みた点で我々のアプローチと似ている。彼らは研究の中で、2015年1月のデータセットから5,535ノードのネットワークについて視覚化したものの、それ以外のデータセットは視覚化が行われず、ネットワーク指標の数値を示すに留まった。彼らは研究の中でダークウェブはコアを持たないフリースケールネットワークの特徴を有すると結論付けた。彼らはダークウェブのグラフ構造の分析に関心を持ち、ネットワークの頑健性について調査を行い、その過程でネットワーク経路長が非常に短いことを発見した。

Griffith ら [23]は、2016年11月から2017年2月にかけてダークウェブのネットワーク構造を調査した結果、クロールによって発見した13,117のドメインから、7,178のアクティブなドメインを見つけた。しかし、彼らはグラフ構造を視覚化しなかった。彼らの研究は、Brunner ら [22]が論文の中で最初の研究であると主張するよりも早くダークウェブのグラフによる分析を行った。彼らはグラフ理論について、社会的な関係の分析と検索可能性などの工学的特性を定量化するために有益であると述べている。彼らはTorプロキシを経由して、通常のウェブと同様の方法でダークウェブのクロールを行った。彼らは収集したデータに対してネットワーク中心性指標の導出、ネットワークの頑健性と脆弱性の分析、相互接続性の分析を行い、ダークウェブ上にはリンクが少ないことを発見した。この結果はDomenico ら [24]の発見と矛盾しない。

Brunner ら [22]は、2018年にダークウェブのネットワーク構造を詳細に分析した。彼らは67,296,302のリンクから34,714のTorドメインを見つけ、約10,000のアクティブなサイトをクロールして収集したデータを分析した。彼らの研究では、収集したデータの10%がダークウェブのネットワーク構造として視覚化された。彼らは独自に開発したクローラーによりTorネットワークのグラフ分析を最初に行い、Torネットワークがウェブと同様にフリースケールネットワークであると結論付けた。彼らの開発したクローラーはTorプロキシを経由してhtmlテキストファイルを収集し、ハイパーリンクを抽出するパーサーとそれを制御するコンダクターモジュール、収集したデータを蓄積するデータベースモジュール、そしてクロールタスクを処理するネットワークモジュールから構成されている。彼らのクローラーはコンテンツの適正を確認するフィルタが組み込まれ、不正なコンテンツを収集対象から自動的に排除する工夫が施されている。彼らはこのフィルタにより、

テキストファイルだけを収集した。彼らは、①取得した順番通りに収集する、②ランダムに収集する、③新しく発見した URL を優先的に収集する、という3つのクローリング戦略によりダークウェブの探索を行った。この方法によりダークウェブから1週間の間で収集されたデータの分析が行われている。彼らは、クローリングにより発見した34,718のアドレスについて、Tor metricsにより推定された70,640と比較して約半数のダークウェブが明らかになったと仮定できると主張している。しかしながら、彼らの収集したアドレスが、その時点で有効なアドレスであったかどうかは判明せず、データが収集された時期についても判然としない。彼らはグラフによる分析を行い、ダークウェブの90%はハブノードから直接到達できると結論付けている。この結果はDomenicoら[24]の発見及びGriffithら[23]の発見と矛盾しない。

Hyperion Grey [29]は、Tor ネットワークの興味深い視覚化を示した。彼らは、2019年に3,700のTor オニオンサービスの視覚化を作成した。彼らの視覚化には、ダークウェブサイトのスクリーンキャプチャをグラフ構造にマッピングする手法が含まれていた。彼らの研究のグラフは、サイト構造の類似性をもとに関係グラフ構造を定義しているため、ハイパーリンク関連のグラフではなかった。

Celleruelo ら [28]は、2020年にTorとI2Pネットワーク間の接続を明らかにするために、ダークウェブネットワークのグラフ構造を作成した。彼らは研究の目的を、違法なコンテンツを特定して法執行機関へ報告するためであるとし、違法行為と戦うためにはTorとI2Pをネットワークグラフで関連付けて調査し発見する必要があると述べている。彼らは、Ahmia [32]およびDUTA-10Kデータセット [12]、ダークウェブからの独自のクローリングによる収集、そしてTorのネットワークサービスディスクリプタから取得した46,562のTorアドレスと2,687のI2Pアドレスのデータを分析した。彼らは様々なデータセットを混合して分析を行ったため、データセットの収集時期についてはあいまいである。彼らの研究にはクローリングしたデータセットのカバレッジについての言及があり、彼らは研究においてカバレッジはTorネットワークの約61%と推定した。彼らの求めたカバレッジはTor Projectによって示されるHidden Serviceドメインの数を基準にした彼らの持つデータセットのドメイン数のカバレッジであった。彼らはHidden Serviceドメインの全てでhttpサーバーが動作しているとは限らない点と、ドメインがいつでもオフラインになる可能性があることを理由に、Torネットワークの正確なカバレッジを推定することは困難であると述べている。彼らは研究の中でグラフによる視覚化を示し、視覚化によって特定のノードを解析することは困難であるとしながらも、I2Pネットワークには明確なサブネットワークを形成するクラスタが存在することを発見した。彼らが研究の中で示したダークウェブネットワークグラフ中心性指標については本研究の結果との比較に役立つ。

ここで挙げた5つの研究は、ダークウェブのネットワーク構造について扱い、本研究の取り扱うテーマと近い。

第3項 ダークウェブのフォーラム構造からグラフを作り分 析する研究

Zhou ら [18]は、2002 年におけるダークウェブから収集されたリンク関係を分析してテロリストグループを特定し、テロリストコミュニティのグラフ視覚化を行った。彼らの関心は米国国内の過激派グループのコミュニティを分析することにあった。彼らは半自動化された手法により、ワールドワイドウェブからクロールを行った。彼らはコミュニティ間の隠された関係を発見する目的でハイパーリンク関係の分析を行い、コミュニティの属性を明らかにする目的でコンテンツの分析を行った。彼らは分析の結果をグラフにより視覚化し、ドメインとコミュニティの関心と関連性が分かり易くなるよう工夫を行った。彼らはグラフによる視覚化を行ったものの、グラフィックアウトの自動化は行わなかった。そして、彼らの調査対象はワールドワイドウェブであり、Tor ネットワークは対象とはしなかった。

Fu ら [19]は、2005 年 4 月から 2006 年 12 月にかけてダークウェブから収集された特定のフォーラムの構造をグラフとして視覚化した。しかしながら、これはダークウェブネットワーク全体をカバーしているわけではない。彼らの関心はフォーラムに特化したクローラーの開発であった。彼らの開発したクローラーは、人間による支援により動作するもので、これによりダークウェブのフォーラムを特定する困難さを解決している。彼らのクローラーはフォーラムを効率的にクロールするための工夫として、言語に依存しない方法を採用し、既に収集したデータを除外する差分クロール、つまり未アクセスページを優先してクロールする方法により、従来のクローラーより効率的に収集できたと結論付けている。

Yang ら [20] [21]は、2006 年と 2007 年にダークウェブフォーラムのクロールを行い、収集されたデータのリンク関係からグラフを作成して視覚化した。彼らの関心はダークウェブフォーラムにおけるソーシャルネットワークグラフの構築と分析にあった。彼らはソーシャルネットワークの視覚化のメリットについて、国家安全保障のために潜在的な脅威を発見し脅威レベルを評価するためのツールと手段を持つことが出来ると述べている。彼らの研究はダークウェブネットワーク全体をカバーしているわけではなかった。彼らの開発したシステムは、ダークウェブフォーラムを対象とするクローラー、収集したデータの分析、ソーシャルネットワークグラフの視覚化、そしてそのインターフェースと外部連携機能により構成されている。彼らの視覚化では、2 次元空間における視覚化効率を最適化する目的でばねモデルによるグラフィックアウトが用いられた。彼らの調査対象はワールドワイドウェブのブログであり、Tor ネットワークは対象としなかった。

ここで挙げた 5 つの研究は、ダークウェブのフォーラム構造をグラフにより分析してい

る。本研究で扱うテーマとは分析観点と対象が異なるものの、グラフ分析と視覚化を扱う研究があり、参考として取り上げた。

第4項 ダークウェブの情報検索にグラフを用いて効率化する研究

Sallaberryら [13]は、ダークウェブでの情報検索のためのグラフの視覚化に取り組んだ。彼らは、検索クエリとページグラフの関係を視覚化した。しかしながら、彼らの視覚化は、ダークウェブ上のハイパーリンクの構造の視覚化ではなかった。彼らは研究の中で情報検索方法の効率化に取り組み、グラフ構造を使って関連する検索結果を見つけやすくした。

ここで挙げた研究は、情報検索の効率化のためにグラフを用いている。本研究で扱うテーマとは分析観点と対象が異なるものの、グラフ分析と視覚化を扱う研究であるため、参考として取り上げた。

第5項 **ダークウェブ犯罪情報の視覚化を目的とする研究**

Park ら [25]は、2020年にダークウェブ上の犯罪情報の視覚化に取り組んだ。彼らはダークウェブに関するいくつかの公開情報を収集し、統計情報を求めて視覚化した。彼らの関心はデータベースによる統計情報の導出であったため、彼らはダークウェブのクロールは行わず、グラフ構造についても視覚化しなかった。

ここで挙げた研究は、ダークウェブにおける犯罪情報の視覚化を目的としている。本研究で扱うテーマであるダークウェブの視覚化という点で一致するため、参考として取り上げた。

第6項 **ダークウェブ測定のガイドラインに関する研究**

Loesing ら [33]は、2010年にダークウェブの統計を分析した。彼らは研究の中で、ダークウェブを測定することは、ダークウェブの利用される国を把握することや匿名通信システムのパフォーマンスを改善させること、そして匿名通信システムの利用目的を調べることに役に立つと述べている。彼らのメンバーはTorの開発者でもあるため、彼らの目的はTorネットワークの匿名性を保護する立場から、Torネットワーク測定に関するガイドラインを示すことであった。彼らはTorシステムの測定に関するガイドラインとして、匿名ネットワークにおける匿名性保護及びプライバシー保護の観点から、必要最低限のデータを取得すること、詳細なデータは公開しないこと、収集するアルゴリズムと取得された統計情報を公開することを求めている。彼らはダークウェブの統計として、Torシステムに固有の測定情報として、Torネットワークに接続するクライアント国情報と、ポートやトラフィックに関する測定を行い、統計情報を示した。彼らの研究はTorシステムにより密接な部分に注力していたため、Torネットワークを構成するTorノードへの言及を行ったものの、Torネットワークを利用してその中に形作られるダークウェブやそのコンテンツの観点からの言及は行わなかった。

ここで挙げた研究は、匿名性を保護する観点からダークウェブの測定を行う場合のガイドラインについて示すことを目的としている。本研究で扱うテーマであるダークウェブの観測の点で参考のため取り上げた。

第7項 ダークウェブのネットワーク構造の時間的変化をグ

ラフにより視覚化する観点における本研究の位置づけ

ダークウェブをクローリングして得られる情報は、ダークウェブの研究にとっての基礎データとなりうるが、悪意を持ってダークウェブを利用したい者にとっても有益な情報であるため、違法行為に対する幫助となりうる可能性の懸念からも、安易に公開されるべき性質のものではない。関連研究を調査した結果、ダークウェブは検索エンジンによる検索では発見できないウェブの領域として位置づけられている [2]ため、本研究においてはTorネットワークに対してウェブと同様にクローリングによるダークウェブネットワークノードの全探索を行った。本研究で収集したデータセットは、ワールドワイドウェブで行われるものと同様の一般的なクローリング手法を用いて、Torネットワークから収集したデータを分析したものである。本研究のクローリングは2018年6月1日から2021年1月30日まで継続的に行われた。既存研究のダークウェブのクローリングでは、数日から数か月程度の短期間の収集が主であり、既存のTorデータセットとしてはDUTA-10K [12]があるが、このデータセットは収集ボリュームに焦点が当てられており、収集された時期については判然としないため、収集時期に焦点を当てる研究には不向きである。また、DUTA-10Kデータセットに含まれる一意のドメイン数は10,367であり、本研究で収集した一意のドメイン数172,740と比べても極めて小さい。さらに、本研究で収集した期間を含む長期間継続的に収集されたダークウェブのデータセットは既存研究には存在しない。さらに、ダークウェブのネットワーク構造の時間的変化をグラフにより視覚化した研究についても見つからなかったことから、本研究ではダークウェブの時系列による視覚化を試みた。また、ダークウェブに対してクローリングを使う研究において、収集したデータからダークウェブネットワークのノードサイズを推定するために標識再捕獲法を使用した研究も見つからなかったことから、標識再捕獲法を使用してダークウェブの大きさを推定した。我々の過去の研究は [30] [31]である。

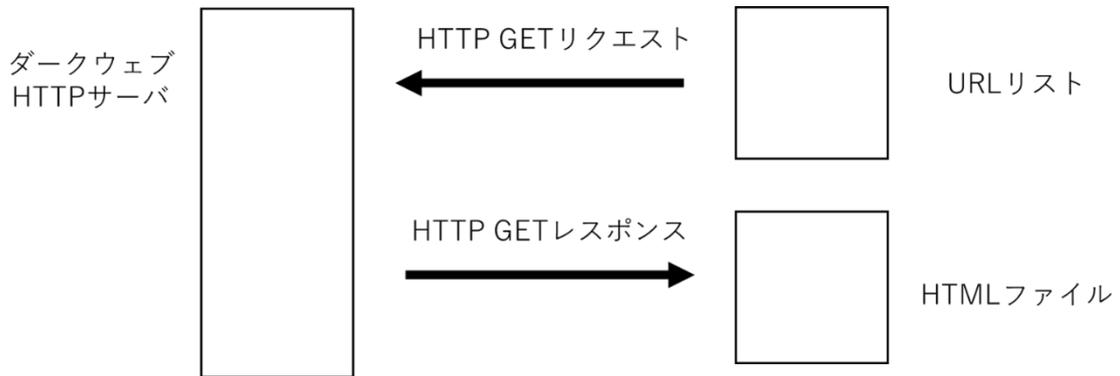
第3節 データセット構築の手法

データセットの構築手法と、スナップショットの生成手法についてそれぞれ示す。本研究で使用するデータは、Tor ネットワークで稼働する http サーバーから、HTTP GET リクエストを発行することで収集される (図表 3)。Tor ネットワークから HTTP GET リクエストにより HTML を取得するコマンドの例を図表 4 に示す。収集対象は、ワールドワイドウェブから効率的にクロールできるとされている幅優先アルゴリズムを使用した。クロール戦略は、Brunner ら [22] によって使用されたアプローチと同様である。本研究では、ネットワークの転送負荷を減らすこととバイナリデータの不要な収集を避けるため、テキストデータのみを選択的にクロールした。テキストデータの選択は、URL 文字列に含まれる拡張子と、HTTP レスポンスヘッダに含まれる Content-Type の値を確認した (図表 5)。また、最終的に取得したデータがテキスト形式でデコードできなかった場合は取得したデータを破棄した。クロウラーの最初のシードには Tor のリンクコレクション² に記載された Tor ページを示す URL を指定した。次に、その Tor ページに含まれるハイパーリンクを抽出した (図表 6)。そして、ハイパーリンクの中からトップレベルドメインが .onion で終わる URL を抽出し、次のクロールターゲットとして蓄積された。クロール用に蓄積されたアドレスはドメイン部分とページ部分に分けられ、それぞれ一様ランダムに選択された。特定のドメインへのアクセスの集中を防ぐために、各リクエストは 10 秒の間隔をあけて同じドメインにアクセスすることを許可した。収集されたデータの鮮度を維持するために、7 日~14 日以内に 1 度も稼働が確認できなかったアドレスをクロール対象の URL リストから削除した。

このようにして収集されたデータセットは、180 日間隔で 6 つのスナップショットに分割した。そして、各スナップショットに含まれる html ファイルから正規表現を使用して .onion リンクのみを抽出し、html ファイルを収集した URL と組み合わせてハイパーリンクグラフデータを生成した。

Loesing ら [33] は Tor システムの測定に関するガイドラインとして、匿名ネットワークにおける匿名性保護及びプライバシー保護の観点から、必要最低限のデータを取得すること、詳細なデータは公開しないこと、収集するアルゴリズムと取得された統計情報を公開することを求めている。本研究の目的を踏まえた上で適法性及び匿名性保護の観点からガイドラインの内容に沿うため、本研究で観測するデータについては、ダークウェブにおいて HTTP リクエストにより取得できるテキスト情報のみを対象とし、Tor システムからの固有の情報や、Tor システムの利用者に関する情報は一切の収集をしていない。本研究の収集に使用するアルゴリズムは公開情報であり、統計情報は本稿で説明される。

² thehiddenwiki.org



図表 3. HTTP GET リクエストを使ったダークウェブからの HTML 情報の取得.

```
curl -x proxy_server:proxy_port -o html_save_path http://target_url
```

図表 4. Tor プロキシを経由して HTTP リクエストを発行するコマンド.

```
1 HTTP/1.1 200 OK
2 Date: Fri, 18 Sep 2020 15:15:37 GMT
3 Server: Apache
4 Expires: Thu, 01 Jan 1970 00:00:00 GMT
5 Pragma: no-cache
6 Content-Type: text/html
```

図表 5. HTTP レスポンスヘッダに含まれる Content-Type の例.

```
grep -o -P "href=¥".*¥" html_save_path | sed -e "s/href=//g" | sed -e "s/¥//g"
```

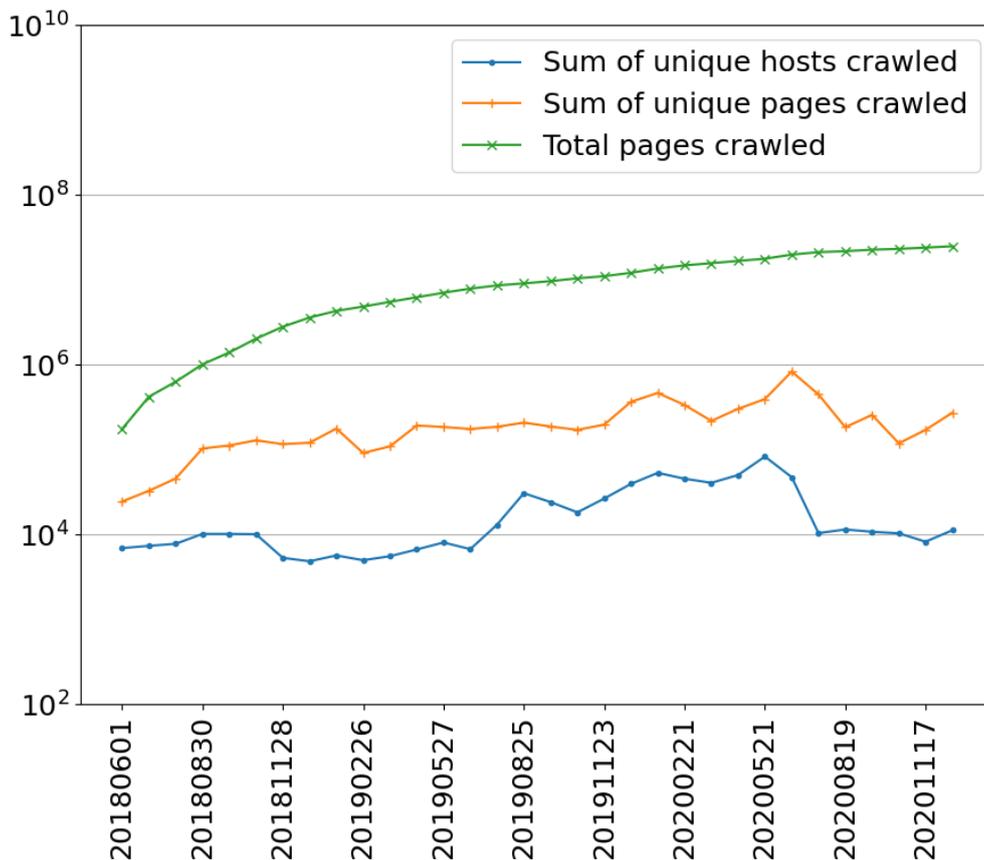
図表 6. HTML ファイルからリンクを抽出するコマンド.

第4節 データセット構築の結果

構築したデータセットと、スナップショットの生成結果についてそれぞれ示す。

第1項 データセットの構築結果

収集は2018年6月1日から2021年1月30日まで繰り返され、全期間中に収集されたページの総数は25,270,157であり、そこから3,763,356の一意のページ、185,989の一意のホスト、そして収集したページに含まれるハイパーリンクを合わせた172,740の一意の.onionドメインが得られた。Torネットワークから収集された一意のページと一意のホストの数および収集したページの総数をそれぞれ30日幅で集計した結果を、図表7に示す。



図表 7. 観測期間中のクローリング数の30日幅のプロット

総収集ページ数積算 (緑), 一意のページ数 (黄), 一意のホスト数 (青)

第2項 スナップショットの生成結果

データセットから時系列分析のためのスナップショットを作成する。スナップショットの日数幅は、データセットを時系列分析する場合における分解能である。クローリング性能の点から、日数幅が小さすぎる場合、ダークウェブの一部しか反映されなくなるため、最小分解能はクローリング性能に影響を受ける。瞬間的なダークウェブの形を知りたい場合は日数幅を短く、ある程度の期間中に起こった変化を全て含めてダークウェブの姿を知りたい場合、必要な日数幅を設定すると良い。本研究ではダークウェブの大局的な変化を理解するための視覚化を目的としているため、スナップショットの数があまりにも多すぎると、論文紙面上では説明しづらい。これらの観点から、本研究ではダークウェブの変化について大局的な変化を見る目的のため、短すぎず長すぎない日数幅として、約半年の180日の日数幅で分割する。

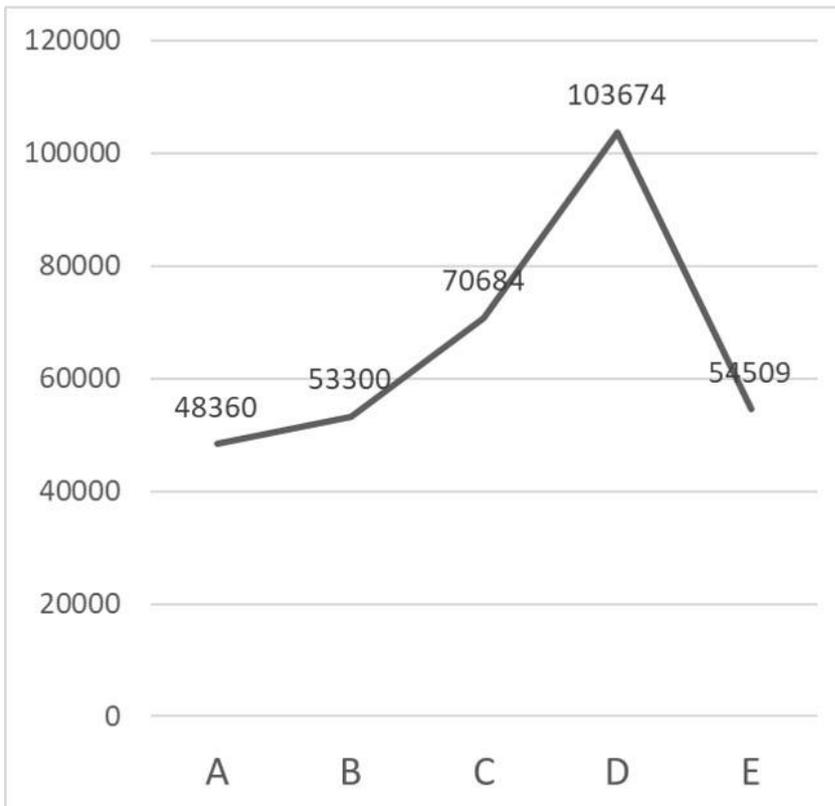
データセットからスナップショットを作成するため、180 収集日単位である図表 8 に示されている期間に従ってデータセットの分割を行った。最後のスナップショット F には、180 収集日に満たないデータが含まれていたため、比較対象からは除外した。次に、正規表現を使用して、各スナップショットに含まれる html ファイルから.onion リンクのみを抽出し、ハイパーリンクグラフデータとした。

図表 8. 各スナップショットの収集期間と収集数

ラベル	クローリング期間	一意のドメイン数
A	2018年6月1日～2018年11月27日	48,360
B	2018年11月28日～2019年5月26日	53,300
C	2019年5月27日～2019年11月22日	70,684
D	2019年11月23日～2020年5月20日	103,674
E	2020年5月21日～2020年11月16日	54,509
F	2020年11月17日～2021年1月30日	17,464
Total	2018年6月1日～2021年1月30日	172,740

第3項 スナップショットの大きさ

5つのスナップショット A~E のノードサイズを比較した(スナップショット F は除外されている)。スナップショット A では、一意のドメインの数は 48,360 だった。スナップショット B から D にかけて着実に増加し、一意のドメインの最大数は 103,674 が記録された。そこからは一転して、スナップショット E では、54,509 ドメインに大幅に減少した(図表 6)。減少した後の値はスナップショット B のドメイン数 53,300 と同程度であった。この結果は、約 1 年間かけて増加したドメイン数と同程度のドメインが 180 日以内に急速に消失したことを示している。



図表 9. スナップショット間の一意のドメイン数

図表 10. スナップショット間の一意のドメイン数

出現したドメイン／消失したドメイン／一致したドメイン

ラベル	スナップショット	消失したドメイン	出現したドメイン	一致したドメイン
Compare-AB	A and B	10,026	14,966	38,334
Compare-BC	B and C	15,084	32,468	38,216
Compare-CD	C and D	21,757	54,747	48,927
Compare-DE	D and E	71,558	22,393	32,116



図表 11. スナップショット間の一意のドメイン数の変化

出現したドメイン／消失したドメイン／一致したドメイン

第5節 データセットの評価

ここでは、収集したデータセットに偏りが無いことを評価するため、データセットから生成されたスナップショット差分の視覚化と、スナップショットにおけるドメインの揮発性、スナップショット間で変化するドメインの偏りについて確認する。更に、データセット全体におけるドメインの生存日数についても確認する。

第1項 スナップショットの差分の視覚化

スナップショット間のドメインの違いを確認するために、スナップショットに含まれるドメインの一覧からなるテキストデータの差分を調べ、失われたドメイン、新しく出現したドメイン、一致したドメインの数をカウントした (図表 7 図表 10)。ドメインの一覧が記録されたテキストデータの差分を調べるために、スナップショットに含まれるドメインの一覧を名前順に辞書順で並べ、Diff ツールにより差分を求めた。Diff ツールは、テキストの内容の違いを計算して表示するテキスト比較ユーティリティであり Unix の世界ではよく使われている。次に、ドメインの増加と減少の偏りを調査するために、Diff ツールの GUI フロントエンドである WinMerge [34] のツールと同様の方法を使用して差分結果を視覚化した。あるスナップショットを次のスナップショットと比較すると、新しいエンタリは赤で表示され、両方に存在するエンタリは白で表示され、存在しなくなったエンタリは灰色で表示される (図表 12)。比較される 2 つのスナップショットの垂直方向の長さは、各スナップショットに含まれる一意のドメインの数によって決定された。長さの違いは、比較されている 2 つのスナップショットを合わせた一意のドメインの総数が違うことを意味している。

第2項 スナップショット間におけるドメインの揮発性

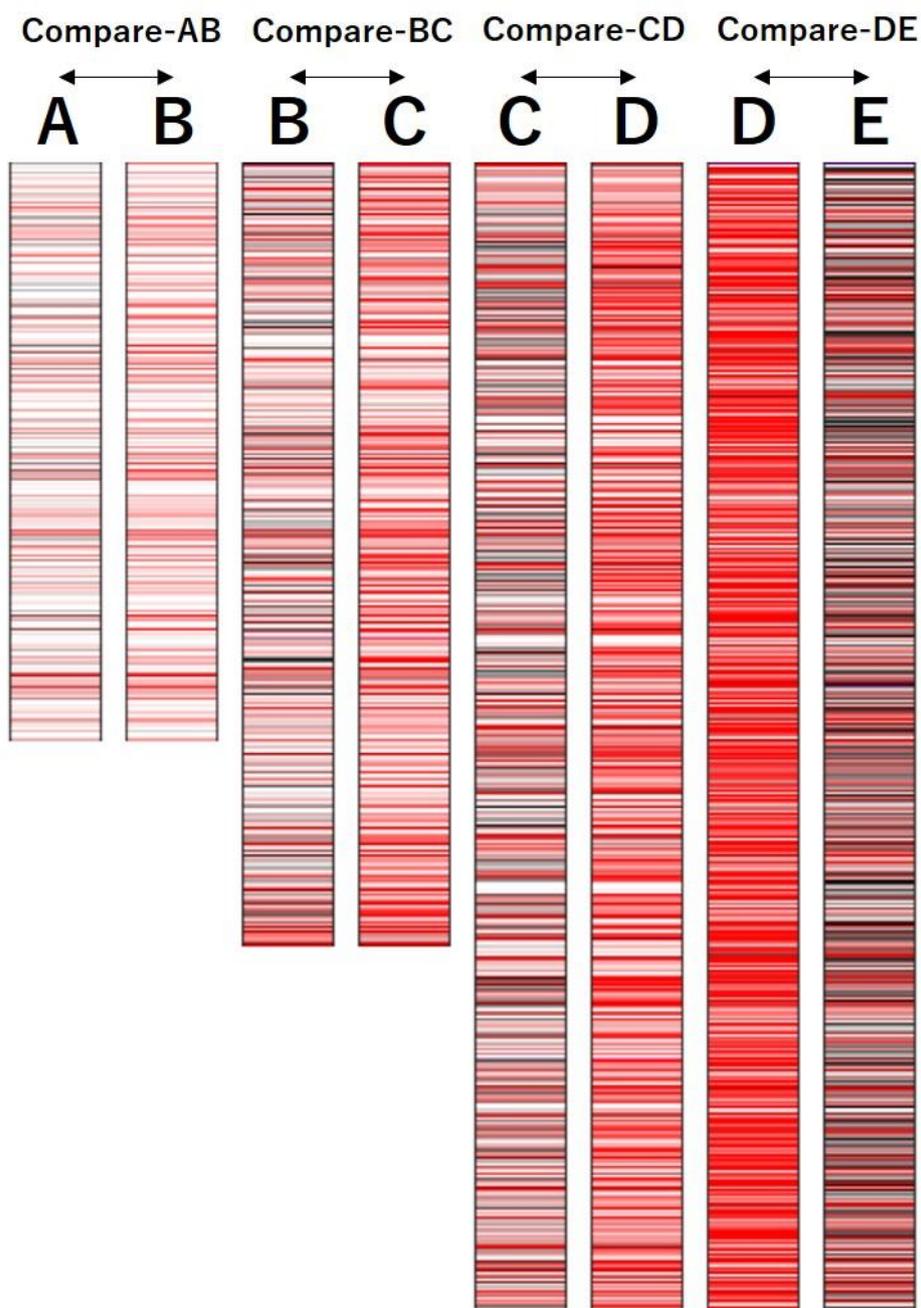
スナップショット A~E については、各スナップショットを次のスナップショットと時系列で比較し、比較をしたスナップショットのそれぞれについて、消失したドメイン、出現したドメイン、一致した一意のドメインをそれぞれカウントした (図表 10)。スナップショット A と B の比較から C と D の比較まで、新しいドメインの数は 14,966 から 54,747 に増加していたが、一致するドメインの数は、A-B と D-E の比較ではそれぞれ 38,334 と 32,116 であり、新しいドメインと同様の増加傾向はみられなかった。(図表 11)。

観測期間中において、一致するドメインの平均数は 39,398 であった。消失したドメインの数は、スナップショット D と E の比較において大幅に増加した。そして出現したドメインの数は減少に転じた。一致したドメインの数も減少したが、出現したドメインや消失したドメインの変化と比較すると、その変化の割合は小さかった。

第3項 スナップショット間で変化するドメインの偏り

出現するドメインと消失するドメインについて、ドメインの偏り具合を調査するために、スナップショット間の差分を視覚化した (図表 12)。スナップショット間の比較において、ドメインが出現もしくは消失している場合、視覚化されたバーではそれぞれ赤または灰色で示される。A から D までのスナップショットの比較では、出現したドメインの数は常に消失したドメインの数を上回った。したがって、これらの比較ペアの右側のバーは主に赤色が支配的となった。スナップショット D と E の比較では、消失したドメインの数が多くなった。したがって、視覚効果として右側のバーでは灰色がより目立つ結果となった。Tor ネットワークにおいて、ドメインは公開鍵暗号により生成されるため、恣意的に鍵を生成しない限り暗号学的な精度のランダム性をもってドメイン名が生成されることから、変化するドメインの数が多くなるにつれ、スナップショット間のドメインテキストの差分の視覚化による変化箇所は均一に分布するようになる。この状態は匿名性の点からは望ましい。一方で、ダークウェブのドメインは総当たりによって意図的に取得することができるため、Tor ネットワークはフィッシングに対して脆弱である。ドメインの偏りは類似したドメイン名を持つ多数のドメインが Tor ネットワーク上に意図的に作成された場合に発生し、Tor ネットワークにおいては、意図した名前アドレスをブルートフォースにより生成するツールが存在する [35]。脆弱な状態では、ダークウェブの多くの部分は人間を騙す目的で類似するドメインが並ぶため、そのようなドメイン群が短期間に登場すると、赤く均一に塗りつぶされた結果が右側のバーに現れる。同様に、短期間で消滅すると灰色に均一に塗りつぶされた領域が左側のバーに現れる。このような状態は匿名性の点からは脆弱である。

この視覚化により得られた結果からは、ドメインの生成と消失にダークウェブの大部分を占めるほどの大きな偏りはみられず、スペクトル結果は概ね均一に分布していることが確認された。



図表 12. スナップショット間のドメインの差分の視覚化

比較した 2 つのスナップショット間における消失したドメインと出現したドメインの差分を視覚化した結果。それぞれの比較において、一意に存在するドメインは赤で表示され、両方に存在するドメインは白で表示され、存在しないエントリは灰色で表示されている。長さの違いは、比較されている 2 つのスナップショットに含まれる一意のドメインの総数の違いを意味する。

次に、2つのスナップショット間のドメインの変化率を計算した。スナップショット A～D の場合、消失した一意のドメインの平均数は、比較対象の2つのスナップショット間に存在した一意のドメインの数の17%であった。しかしながら、スナップショット D～E の比較では、消失した一意のドメインの割合は57%であった。スナップショット間で一致した一意のドメインの割合は、スナップショット A-B から D-E で60%から25%に低下した。

第4項 ドメインの生存日数

ドメインの生存日数を調べるために、ドメインごとに最後に観測された日と最初に観測された日の差分を求め、生存日数とした。

$$dt_{survive} = dt_{lastseen} - dt_{firstseen} + 1 \quad (1)$$

ここで、 $dt_{survive}$ は生存日数を表し、 $dt_{lastseen}$ は最後に観測された日を表し、 $dt_{firstseen}$ は最後に観測された日を表す。

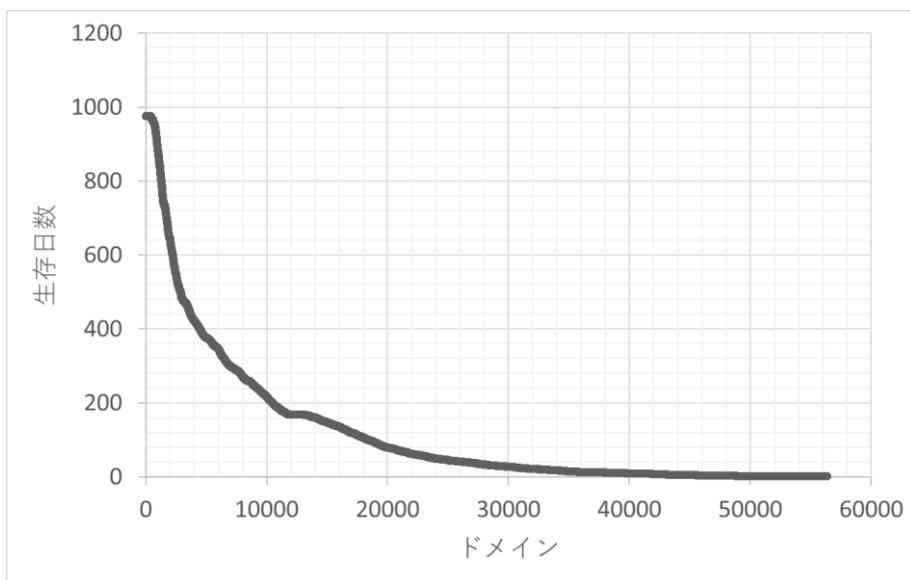
データセットに含まれる全てのドメインに対して生存日数を求め、中央値、平均、分散、最大値、最小値をそれぞれ求めた。データセットの収集期間から、生存日数の最大値は975日である。データセットは、期間中に1回以上の観測が行われたデータで構成されるため、生存日数の最小値は1日である。

全てのスナップショットに含まれるドメインの中から、生存が確認できたドメインだけを対象に、生存日数の中央値、平均、分散、標準偏差、最大値、最小値をそれぞれ求めた(図表13)。標本数 n は56,336であった。生存日数は、最後に観測された日と最初に観測された日の差分である。結果から、観測されたダークウェブの多くはおおよそ1か月で消滅することがわかる。これはMiyataら[36]の調査による2001年のウェブの生存日数である1,108日と比べると著しく短い。

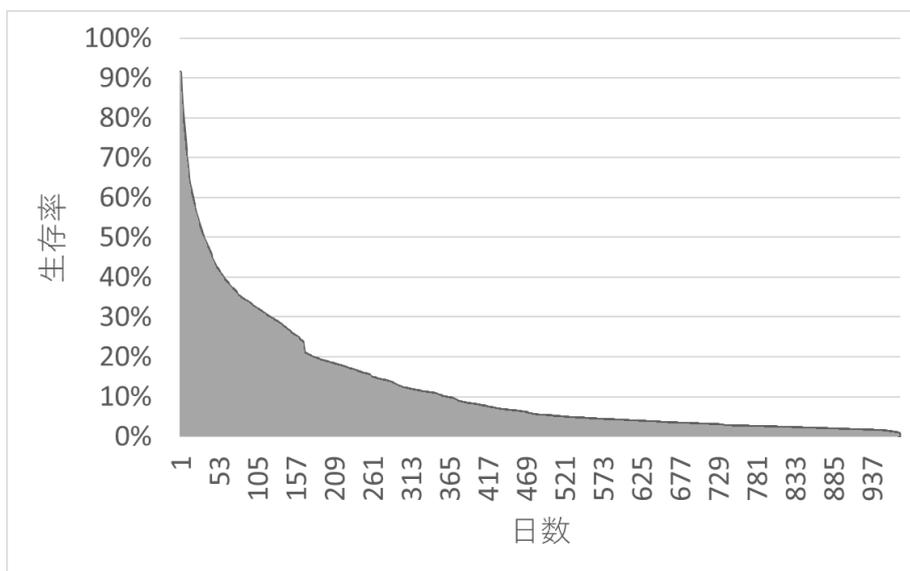
図表 13 データセットに含まれる全てのドメインの生存日数表

項目	日数
生存日数の中央値	33
生存日数の平均	119
生存日数の分散	36,304
生存日数の標準偏差	191
生存日数の最大値	975
生存日数の最小値	1
標本数 (n)	56,336

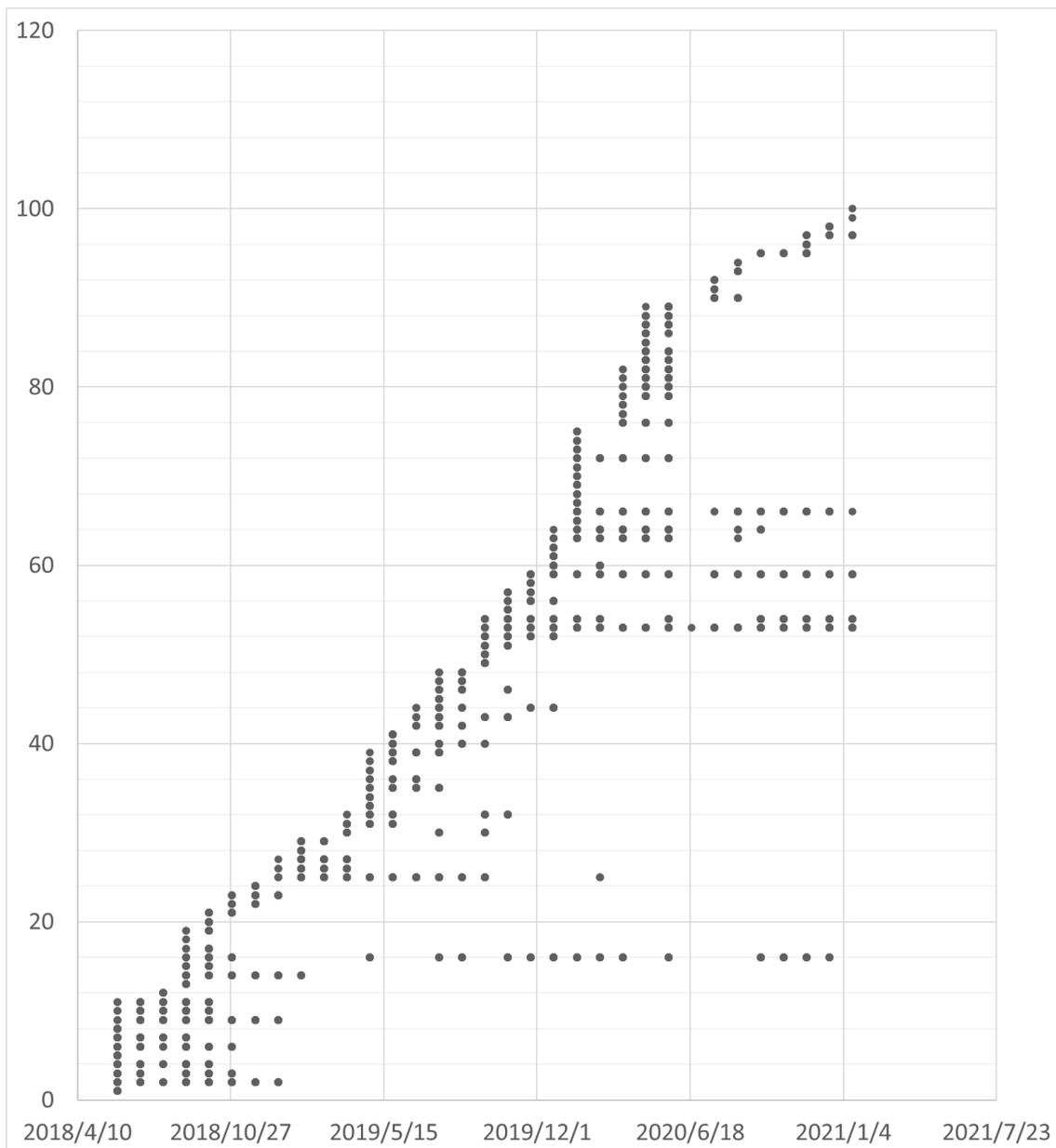
データセットに含まれる全てのドメインの生存日数をプロットしたものを図表 14 に、ドメインの生存率を割合でプロットしたものを図表 15 に示す。1995 年から 1999 年のウェブを対象にした Spinellis [37] の 2003 年の調査によるとウェブサイトの寿命の半減期はおおよそ 4 年であり、1998 年から続く技術系記事を対象にした Iio [38] の 2011 年の調査によると半減期はおおよそ 6 年とされる。データセットにおける半減期は中央値と等しく 33 日であることから、ダークウェブの半減期についてもウェブと比べて 1.5%~2.3% しかなく、著しく短い。データセットから無作為にサンプリングした 100 ドメインの活性状態を図表 16 に示す。縦軸はドメインに割り振られた一意の番号を表し、30 日の間に 1 度でも観測されるとプロットされる。図からは、大多数のドメインが短期間で観測されなくなる様子が確認できる。



図表 14 データセットに含まれる全てのドメインの生存日数のプロット



図表 15 データセットに含まれる全てのドメインの生存率



図表 16 データセットから無作為にサンプリングした 100 ドメインの活性状態

30 日ごとの集計。縦軸はドメインを表し、横軸は時間である。長期間観測されるドメインは横に長く継続してプロットされる。

第6節 議論

ダークウェブの観測を開始してから 2 年の間、ダークウェブネットワークのノード数は増え続け、2018 年 6 月から 2020 年 5 月にかけて、ダークウェブはより多様で複雑になった。2019 年、法執行機関はダークウェブ上の多数の違法サイトを削除することに成功した。これらのサイバー犯罪対策にもかかわらず、ダークウェブのサイズに大きな影響は与えられなかった。いくつかの要因がダークウェブの成長を促したと考えられる。最も重要な要因の 1 つは、犯罪ビジネスが確立され、その取引規模が大きくなったことである。

本研究の結果によると、2020 年 5 月以降、ノードの数は 180 日間で大幅に減少し、過去 2 年間に経験した増加と同じであった (図表 46)。2020 年 9 月 22 日、FBI はダークウェブで 179 人の麻薬密売人を逮捕した [39]。2021 年 1 月 12 日、世界最大のダークウェブマーケットプレイスとして知られるダークマーケットがユーロポール [40]によって削除された。これらは犯罪に使われるダークウェブを小さくすることに成功したのだろうか？ Tor プロジェクトの記録によると、これらの取り組みは成功していない結果を示している。すなわち、ダークウェブは小さくなってはいない。図表 17 は、Tor プロジェクトによって測定された V2 アドレスの数を示している。V2 アドレスとは、Tor のバージョン 2 で使用される古い形式のアドレスである。同期間にアクティブアドレスの数が約 100,000 から 200,000 を超えるまで急激に増加し、直後に約 175,000 に減少したことが記録されている。彼らのグラフによると、急激な増加とその後の減少があったものの、アクティブなアドレスの数は増加前の数まで減少していないようである。各スナップショットで観察されたドメイン数の比較から、一致するドメインの平均数は 39,398 で一定であった (図表 10)。本研究の観測データは、ダークウェブ内の揮発性の高いノードの割合が 40 % から 75 % に増加した可能性があることを示唆している (図表 18)。我々の過去の研究 [30]は、観察されたドメインのスナップショットに差分を適用することにより、ダークウェブドメインの名前空間における出現や消失の分布に偏りが無いことを示した。この結果からは、ダークウェブを故意に操る恣意的な存在の可能性を否定し、ダークウェブでは揮発性の高いノードの割合が徐々に増加している可能性についての示唆が得られた。

ここからはいくつかの関連研究との差異について議論する。Broder ら [15]は、1999 年 5 月から 1999 年 10 月にかけてワールドワイドウェブをクローリングして 200 万ページ以上のリンクを辿り、抽出された 15 億のリンクを使用してウェブのグラフを作った。本研究では 2018 年 6 月 1 日から 2021 年 1 月 30 日にかけてダークウェブをクローリングして 2,500 万ページ以上のリンクを辿り、抽出された 17 万のリンクを使用してグラフを作った。本研究の収集期間は Broder らより短いものの、収集したページ総数は多く、抽出したリンク数は少ない点が異なる。彼らは測定したウェブ上のノード数を報告したが、ウェブのノード数の推定は行わなかった。我々は測定したダークウェブのノード数を報告し、ノード

数の推定値を示した点が異なる。

Toyoda ら [14]は、1999 年から 2003 年にかけて日本のウェブを対象としてクローリングを行い、それらの時系列グラフを分析し、視覚化した。彼らの研究では、ウェブの個々の詳細な変化について焦点を当てて分析がなされ、本研究ではダークウェブ全体について大局的な変化について扱う点が異なる。

Heer ら [27]は、ソーシャルネットワークグラフの構造を視覚化した。本研究ではダークウェブを対象としている点が異なる。

彼らの研究は、いずれもダークウェブについては視覚化していないため、ダークウェブについても同様の観点から研究を進めることは有益である。

Domenico ら [24]は、2013 年から 2015 年の期間に Tor ネットワークの内部からサンプリングしたサーキットの繋がりから 3 つのグラフを生成して分析を行った。本研究では 2018 年から 2021 年にかけてクローリングにより収集したデータセットからリンク関係を示す 5 つのグラフを生成して分析した。彼らの研究は、複数のデータセットの比較を試みた点で我々のアプローチと似ているが、彼らの分析したグラフは Tor ネットワーク内部のサーキットのグラフであり、ダークウェブのリンク関係を示すグラフではない。更に、彼らはグラフ構造を詳細に分析し、ネットワークの平均次数は次第に低下していることを示したが、ノードの生存日数について分析をしなかった点が本研究と異なる。

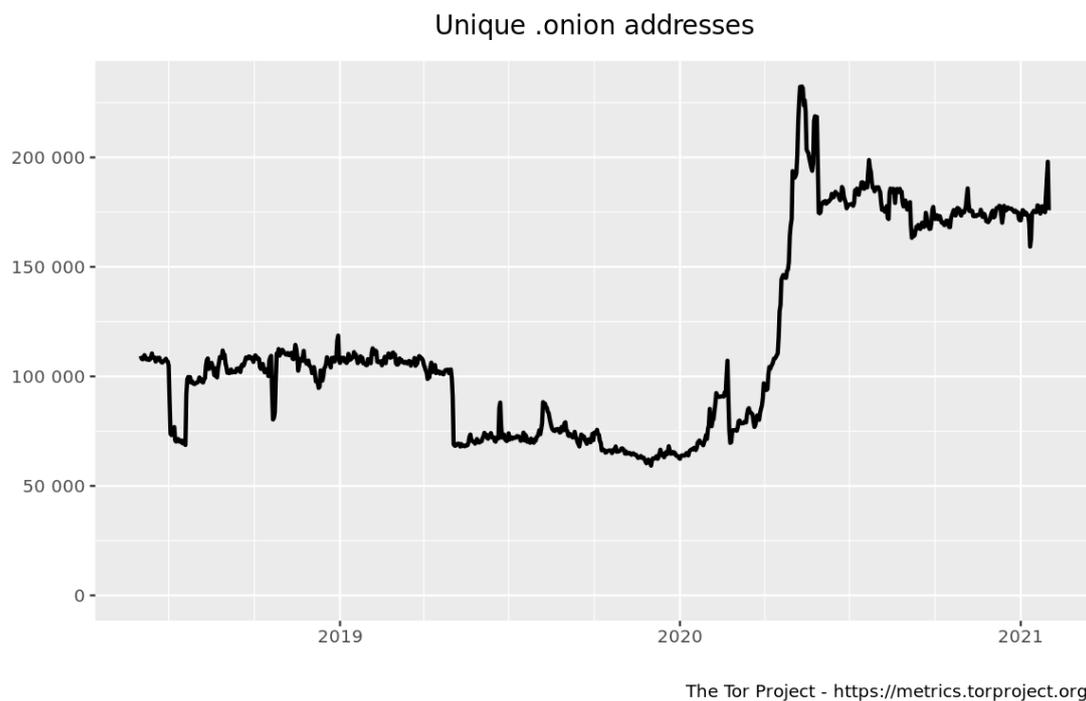
Griffith ら [23]は、2016 年から 2017 年にかけてダークウェブからクロールによって発見した 13,117 のドメインを分析した。本研究では 2018 年から 2021 年にかけてダークウェブをクローリングして発見した 17 万のドメインを分析した。彼らの分析は時系列の視点で行っていない点が本研究と異なる。彼らは分析によりネットワーク中心性指標を求め、次数中心性の高い上位ノードを示した点は我々と同様である。本研究の測定結果の中で彼らの分析に最も近いスナップショット A の出力次数（図表 61）を比較すると、リンクコレクション等の一部の結果は一致がみられるものの、多くの部分は異なる結果であった。この結果の違いは、分析対象のドメイン数がおおよそ 10 倍以上の差となって表れている両者におけるクローリング範囲の違いがあることと、更には本研究で求めたダークウェブの生存日数は 33 日であったことから、1 年の間にダークウェブの中身が大きく入れ替わった可能性などが考えられる。

Brunner ら [22]は、2018 年にダークウェブからおおよそ 1 週間の内にクロールによって収集した 34,714 の Tor ドメインを分析した。本研究では 2018 年から 2021 年にかけてダークウェブを継続的にクローリングして発見した 17 万の Tor ドメインを分析した。彼らは独自に開発したクローラーにより収集したデータからダークウェブのリンク関係のグラフを作った点で本研究とよく似ているが、収集期間は本研究のもの比べると極めて短い。彼らの分析は時系列の視点で行っていない点が本研究と異なる。更に、彼らの研究はクローラーの作成と収集したコンテンツの統計分析を重視していたが、本研究では視覚化による時系列分析によりダークウェブの変化を理解することを重視している点も異なる。

彼らの研究ではダークウェブの入力次数中心性の上位 10 件が示されていたため、本研究の測定結果の中で彼らの結果と最も近いスナップショット A 及びスナップショット B の入力次数 (図表 54, 図表 55) と比較すると、その時期に人気のあるマーケットやホスティングサービス、匿名メールサービス、リンクコレクションなどがそれぞれの結果から確認でき、おおよそ近い結果であった。

Cilleruelo ら [28] は、2020 年にダークウェブからのクローリングを含むいくつかのデータセットを混合し、46,562 の Tor アドレスと 2,687 の I2P アドレスをもとにダークウェブのグラフによる分析を行った。彼らの分析は時系列の視点で行っていない点が本研究と異なる。彼らは様々なデータセットを混合して分析を行ったため、データセット全体の収集時期についてはあいまいである。彼らは研究の中でダークウェブネットワークグラフ中心性指標の上位ノードを示していたことから、本研究の測定結果の全てのデータから求めた中心性指標であるスナップショット A-F の入力次数 (図表 60) 及びスナップショット A-F の出力次数 (図表 67) とそれぞれ比較したところ、入力次数と出力次数のいずれも一致が無かった。この結果は、Griffith ら [23] との比較と同様に、クローリング時期や範囲が異なることで、ダークウェブが短期間の内に大きく入れ替わってしまう影響が大きいものと考えられる。

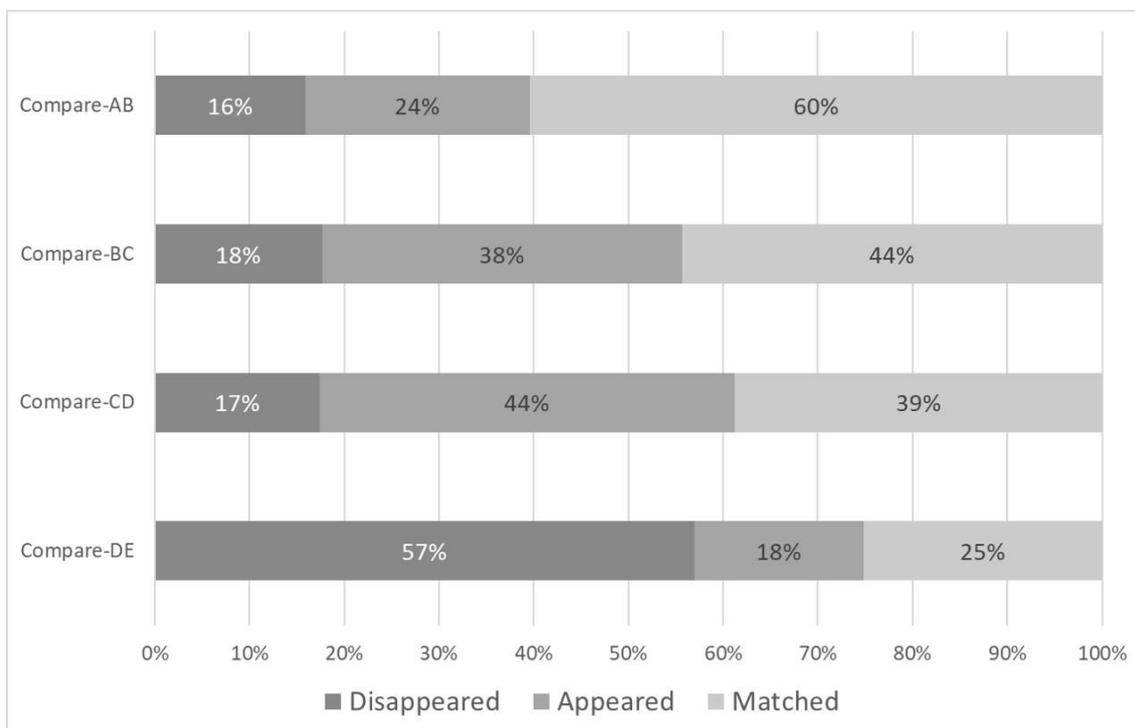
いくつかの関連研究と本研究との比較を行った結果、ダークウェブの分析結果は比較的短期間の内に大きく異なる結果に変化してしまうと考えられ、本研究で求めた生存日数の統計値 (図表 13) が示す通り、揮発性の高いネットワークであることが示唆される結果となった。



図表 17. オニオンサービスの数

オニオンサービスの数 (v2 address only) from June 1, 2018, to January 31, 2021.

Source: metrics.torproject.org (under Creative Commons Attribution 3.0 License - CC BY 3.0 US)



図表 18. スナップショット間で変化のあるドメインの割合
出現したドメイン／消失したドメイン／一致したドメイン

第7節 本章のまとめ

本章では、ダークウェブの探索と、データセットの構築、スナップショットの作成について記載した。観測時点においてダークウェブとして人気のある Tor ネットワークを対象に観測を行った。ダークウェブは、その言葉の登場にあたり、検索エンジンによる検索では発見できないウェブの領域として位置づけられているため、自身により観測手段を確保する必要がある。本研究では、ダークウェブは html により繋がるネットワークである点に着目して、ウェブの真の目的は人間に対する情報とコミュニティへの誘導の提供であるとの仮説の元、ワールドワイドウェブと同様の手法でダークウェブの観測を行った。つまり、人間がダークウェブにアクセスしてコミュニティを追いかけ続けることが出来るのであれば、同様の方法で機械的に追い続けることが出来るというウェブの設計思想がダークウェブにおいても同様に働くであろうとの仮説に基づく。この仮説を証明することは困難であるが、本研究の観測においてはこの仮説は有効に機能し続けた。つまり、クローリングによるダークウェブ探索の動作が途中で途切れ、観測が停止することはなかった。本研究における観測により、おおよそ 2 年に渡る 2018 年 6 月 1 日から 2021 年 1 月 30 日までの期

間において、25,270,157 ページの html テキストファイルを収集した。これらのテキストファイルは、幅優先探索により Tor ネットワークから収集された。これらのテキストファイルからハイパーリンクを抽出し、グラフデータを構築した。そして、データセットは180日の期間により時系列ごとに分割し、分析のためのダークウェブのスナップショットとした。これらのスナップショット間において、ドメインテキストの差分の視覚化を行い、ダークウェブのドメイン名の分布や生成・消失について偏りがみられないことを確認した。更に、ダークウェブのドメインの生存期間を調べ、統計情報を提供した。データセット全体からドメインの生存期間について分析を行い、生存が確認できたドメインだけを対象に、生存日数の中央値、平均、分散、標準偏差、最大値、最小値をそれぞれ求めた結果、観測されたダークウェブの多くはおおよそ1か月で消滅することがわかり、Miyataら [36]の調査による2001年のウェブの生存日数である1,108日と比べると著しく短いことが明らかとなった。

本章の結果として、ダークウェブのネットワークはドメイン名の観点から偏りは確認できなかった。そして、ダークウェブにおけるドメインの生存期間から求まるダークウェブの半減期についてもワールドワイドウェブと比べて1.5%~2.3%しかなく、著しく短いことが示された。

第4章

ダークウェブネットワークの視覚化

第1節 概要

第3章で作成したスナップショットの有向グラフデータについてばねモデルを用いて視覚化する。最初にダークウェブネットワークの視覚化技術についての関連研究を示し、その比較を通して本研究が提案する視覚化の方法を説明し、視覚化により得られた結果を示す。次に、視覚的に抽出した標本の特徴評価を行い、ネットワークグラフ中心性指標の導出、上位10位のハブノードの抽出をそれぞれ求めた上で、グラフによる視覚化結果と中心性指標との比較を行い、ダークウェブネットワークの時系列的な変化を追う。

第2節 関連研究

ここでは、ダークウェブの時系列ハイパーリンクグラフ構造を視覚化し、ネットワーク中心性指標と比較を行うため、ダークウェブに関する視覚化を行う研究を中心に関連研究を示す。

第1項 ダークウェブの構造からグラフを作り分析する研究

Brunnerら [22]は、2018年にダークウェブのネットワーク構造を詳細に分析した。彼らはTorドメインから約10,000のアクティブなサイトをクロールして収集したデータを分析した。彼らの研究では、収集したデータの10%がダークウェブのネットワーク構造として視覚化された。彼らはグラフによる分析を行い、ダークウェブの90%はハブノードから直接到達できると結論付けている。この結果はDomenicoら [24]の発見及びGriffithら [23]の発見と矛盾しない。

Hyperion Grey [29]は、Tor ネットワークの興味深い視覚化を示した。彼らは、2019 年に 3,700 の Tor オニオンサービスの視覚化を作成した。彼らの視覚化には、ダークウェブサイトのスクリーンキャプチャをグラフ構造にマッピングする手法が含まれていた。彼らの研究のグラフは、サイト構造の類似性をもとに関係グラフ構造を定義しているため、ハイパーリンク関連のグラフではなかった。

Celleruelo ら [28]は、2020 年に Tor と I2P ネットワーク間の接続を明らかにするために、ダークウェブネットワークのグラフ構造を作成した。彼らは、Ahmia データセット [32] および DUTA-10K データセット [12]、ダークウェブからの独自のクローリングによる収集、そして Tor のネットワークサービスディスクリプタから取得した 46,562 の Tor アドレスと 2,687 の I2P アドレスのデータを分析した。彼らは様々なデータセットを混合して分析を行ったため、データセットの収集時期についてはあいまいである。彼らは研究の中でグラフによる視覚化を示し、視覚化によって特定のノードを解析することは困難であるとしながらも、I2P ネットワークには明確なサブネットワークを形成するクラスタが存在することを発見した。彼らが研究の中で示したダークウェブネットワークグラフ中心性指標については本研究の結果との比較に役立つ。

ここで挙げた研究は、ダークウェブのネットワーク構造について視覚化を行うため、本章で取り扱うテーマと近い。

第2項 ダークウェブのフォーラム構造からグラフを作り分 析する研究

Zhou ら [18]は、2002 年におけるダークウェブから収集されたリンク関係を分析してテロリストグループを特定し、テロリストコミュニティのグラフ視覚化を行った。彼らの関心は米国国内の過激派グループのコミュニティを分析することにあつた。彼らは半自動化された手法により、ワールドワイドウェブからクローリングを行った。彼らはコミュニティ間の隠された関係を発見する目的でハイパーリンク関係の分析を行い、コミュニティの属性を明らかにする目的でコンテンツの分析を行った。彼らは分析の結果をグラフにより視覚化し、ドメインとコミュニティの関心と関連性が分かり易くなるよう工夫を行った。彼らはグラフによる視覚化を行ったものの、グラフィレイアウトの自動化は行わなかった。そして、彼らの調査対象はワールドワイドウェブであり、Tor ネットワークは対象とはしなかった。

Fu ら [19]は、2005 年 4 月から 2006 年 12 月にかけてダークウェブから収集された特定のフォーラムの構造をグラフとして視覚化した。しかしながら、これはダークウェブネットワーク全体をカバーしているわけではない。彼らの関心はフォーラムに特化したクロー

ラーの開発であった。

Yang ら [20] [21]は、2006年と2007年にダークウェブフォーラムのクローリングを行い、収集されたデータのリンク関係からグラフを作成して視覚化した。彼らはソーシャルネットワークの視覚化のメリットについて、国家安全保障のために潜在的な脅威を発見し脅威レベルを評価するためのツールと手段を持つことが出来ると述べている。彼らの研究はダークウェブネットワーク全体をカバーしているわけではなかった。彼らの視覚化では、2次元空間における視覚化効率を最適化する目的でばねモデルによるグラフィックレイアウトが用いられた。彼らの調査対象はワールドワイドウェブのブログであり、Tor ネットワークは対象としなかった。

ここで挙げた5つの研究は、ダークウェブのフォーラム構造をグラフにより分析している。本章のテーマとは分析観点と対象が異なるものの、グラフ分析と視覚化を扱う研究があり、参考として取り上げた。

第3項 **ダークウェブの情報検索にグラフを用いて効率化する研究**

Sallaberry ら [13]は、ダークウェブでの情報検索のためのグラフの視覚化に取り組んだ。彼らは、検索クエリとページグラフの関係を視覚化した。しかしながら、彼らの視覚化は、ダークウェブ上のハイパーリンクの構造の視覚化ではなかった。彼らは研究の中で情報検索方法の効率化に取り組み、グラフ構造を使って関連する検索結果を見つけやすくした。

ここで挙げた研究は、情報検索の効率化のためにグラフを用いている。本章のテーマとは分析観点と対象が異なるものの、グラフ分析と視覚化を扱う研究であるため、参考として取り上げた。

第4項 **ダークウェブ犯罪情報の視覚化を目的とする研究**

Park ら [25]は、2020年にダークウェブ上の犯罪情報の視覚化に取り組んだ。彼らはダークウェブに関するいくつかの公開情報を収集し、統計情報を求めて視覚化した。彼らの関心はデータベースによる統計情報の導出であったため、彼らはダークウェブのクローリングは行わず、グラフ構造についても視覚化しなかった。

ここで挙げた研究は、ダークウェブにおける犯罪情報の視覚化を目的としている。本章のテーマであるダークウェブの視覚化という点で一致するため、参考として取り上げた。

第5項 本研究の位置づけ

本研究は、ダークウェブネットワークの変化を視覚化することによって、ダークウェブネットワークの変化と変遷についての理解を深める手段を提供することを目的としている。我々の知る限り、ダークウェブネットワークは巨大で複雑で流動的であるため、ダークウェブの状態を理解することは簡単ではない。ダークウェブの全てを瞬時にキャプチャすることは困難であり、ダークウェブの全てを一人の人間が追いかけて理解することは困難である。これは既存の研究で言及 [28]されるものと矛盾しない。

視覚化は、大規模で複雑なネットワークがどのように変化してきたかという問題に対して直感的な理解を提供するが、既存の研究では詳細な分析や数値による視覚化に留まり、ダークウェブの変化という視点で視覚化を活用していなかった。そのため、本章では、ダークウェブネットワークを時系列グラフにより視覚化し、ダークウェブの変化から様々な洞察を得るための手段を提供する。そして、時系列変化とグラフ中心性指標を比較し、視覚化することの意義を示す。

第3節 グラフによる視覚化の手法

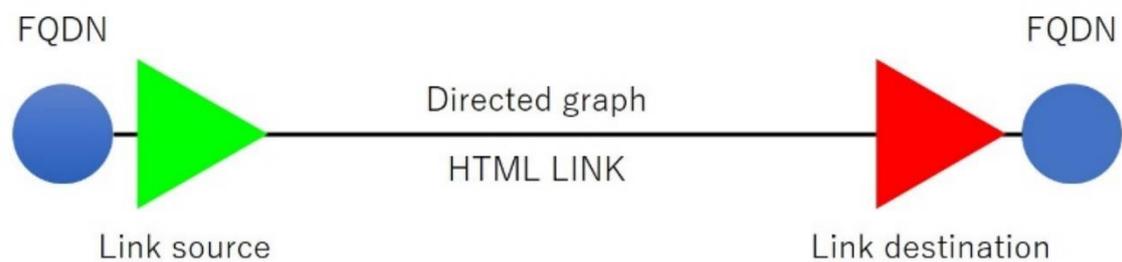
本研究のテーマであるダークウェブハイパーリンクグラフの視覚化手法について解説する。視覚化のため、スナップショットから時系列グラフを用意する。スナップショットからなる時系列グラフは次の式で定義される。

$$\{G_t := (V_t, E_t) | 1 \leq t \leq T\} \quad (1)$$

ここで、 V_t はドメインノードのリスト、 E_t はノード V_t の関係性を表し、ドメイン間のハイパーリンク関係を記述する関係データである。これは、記録された入力ノードと出力ノードを持つテーブルデータで表される。スナップショット A から E、すなわち $T=5$ から 5 つのグラフ $G_{1,2,3,4,5}$ を作成した。グラフ内のノード数を削減するために、URL からドメインを抽出し、ドメイン間のハイパーリンク関係を示すグラフデータを作成した。

グラフデータは、自動的なレイアウトのためにばねモデル [41]で視覚化した。ばねモデルを使用すると、他のグラフレイアウトよりも全体像をコンパクトに描画することが出来るため、全体の形をとらえやすい。ばねモデルの実装はオープンソースで人気のある D3 力学有向グラフレイアウト [42]を使用した。同様の機能を持つものであれば、他のグラフ描画ツールを使用しても良い。参考のため、ばねモデル [41]、円形レイアウト [43]、階層型レイアウト [44]で視覚化した場合の違いを付録 A 図表 50, 図表 51, 図表 52 及び図表 53 にそれぞれ示す。その他のグラフ描画方法については Herman らの調査 [45]が参考になる。

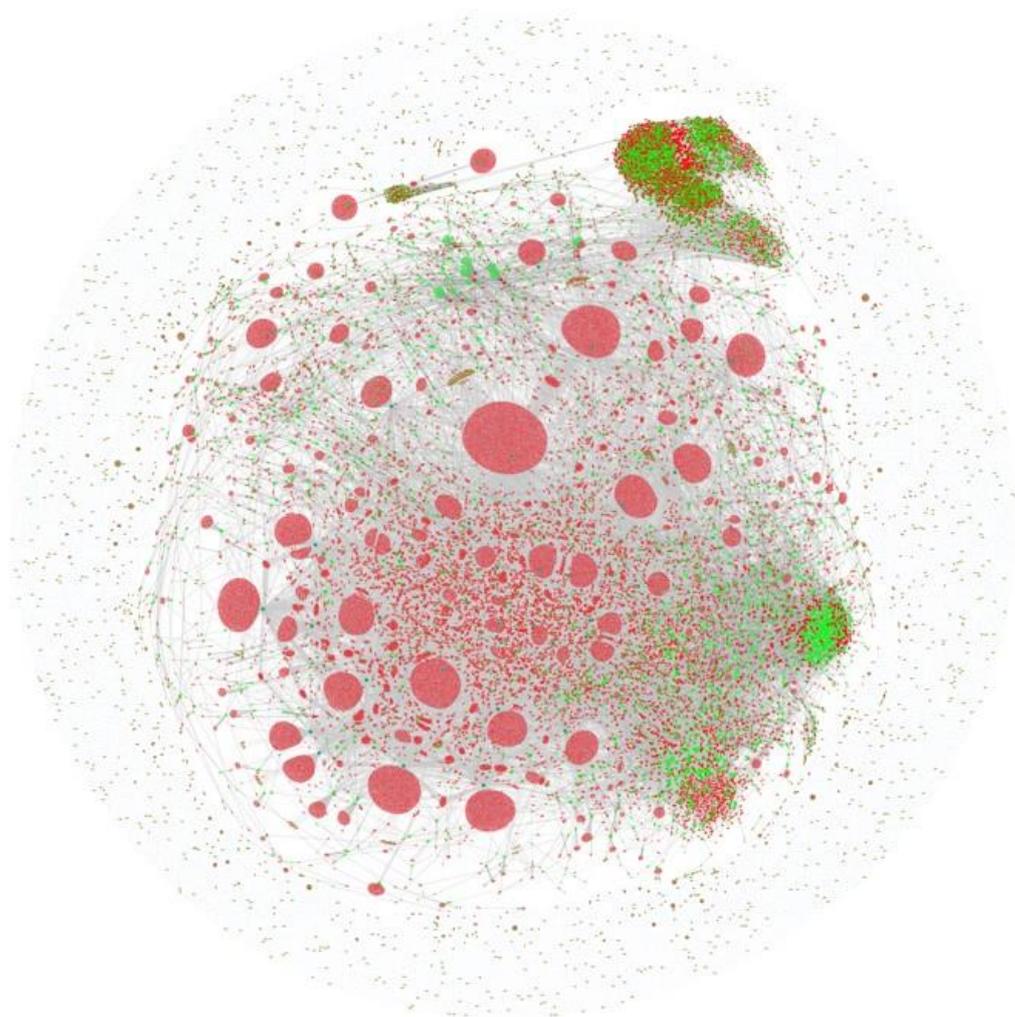
次に、エッジの方向が視覚的に表現されるようにグラフの色を調整した。ハイパーリンクのソースは緑でマークされ、ハイパーリンクのデスティネーションは赤でマークされる (図表 19)。ノードの色は青でマークされる。描画順序はエッジごとにハイパーリンクのソース (緑)、ハイパーリンクのデスティネーション (赤) の順でエッジの数だけ繰り返して描画し、ノード (青) を最後に描画した。エッジの大きさはばねモデル、クーロン力、重力、および衝突検出が行われるようそれぞれ D3 パラメータに設定する。ばねモデルを使用した自動グラフレイアウトには、スナップショット内のすべてのノードと、5 % のランダムにサンプリングされたエッジを使用した。



図表 19. ハイパーリンク関係の方向性の色による視覚化

第4節 グラフによる視覚化の結果

スナップショット A から E の 5 つについて視覚化した結果を，図表 21，図表 22，図表 23，図表 24，図表 25 にそれぞれ示す．データセット全体の視覚化結果は図表 20 である．

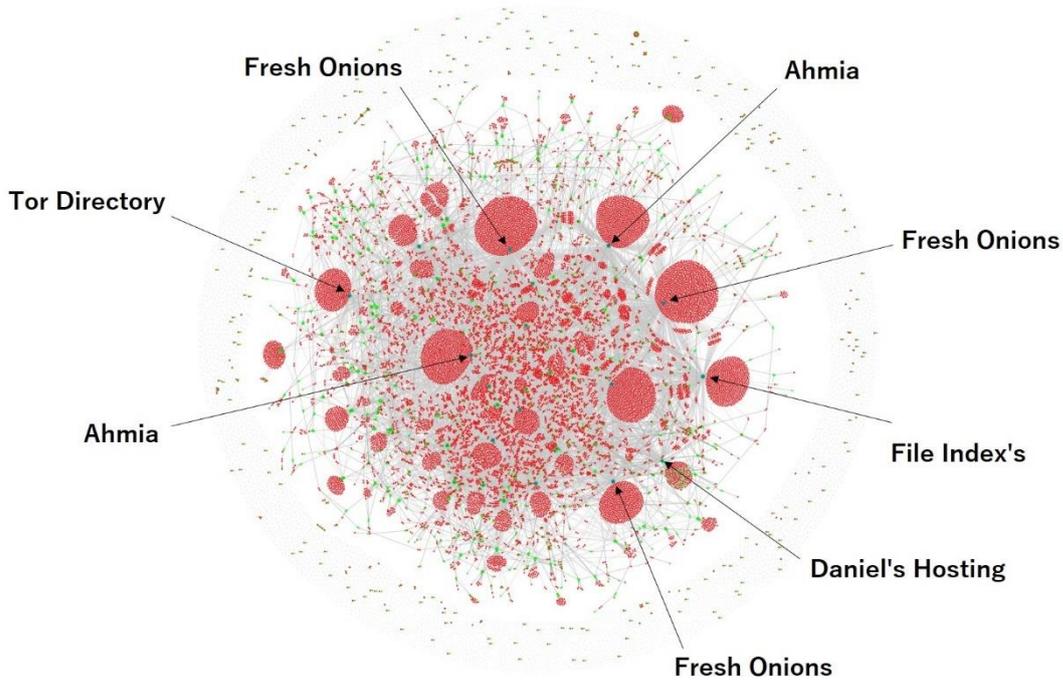


図表 20. ダークウェブハイパーリンクネットワークグラフの視覚化

期間は 2018 年 6 月 1 日から 2021 年 1 月 30 日まで

第1項 スナップショットAの視覚的標本の抽出

スナップショット A の視覚化結果を図表 21 に示す。この視覚化では、約 8 つのノードが視覚的に目立つ位置で広い領域をもつため、これらをスナップショット A から視覚的に抽出した標本（付録 図表 96 から図表 103 に掲載）とした。これらの標本を調べると、リンクコレクション（図表 96, 図表 98, 図表 99, 図表 101, 図表 103）、検索エンジン（図表 97, 図表 102）、ホスティングサービス（図表 100）であった。グラフでは、赤色は宛先ハイパーリンクを表し、緑色はハイパーリンクソースを表している。スナップショット A による視覚化からは、ハブノードが多くの宛先ノードにリンクされている様子が示されている。

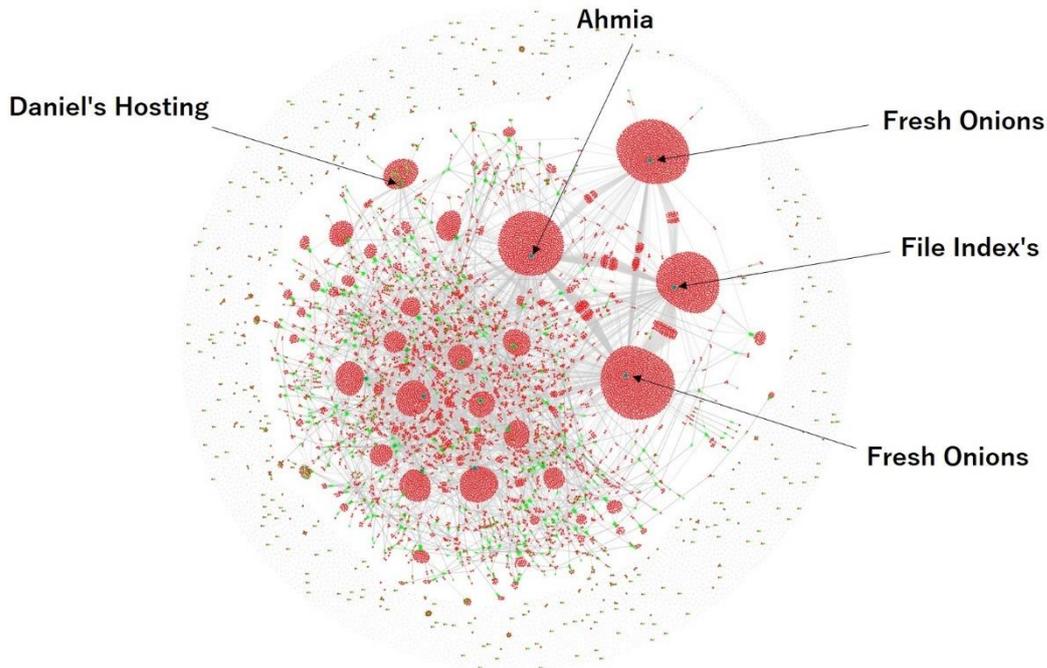


図表 21. スナップショット A のハイパーリンクネットワークグラフの視覚化

Ahmia, Fresh Onions, File Index's, Daniel's Hosting Tor Directory はそれぞれ固有名詞

第2項 スナップショットBの視覚的標本の抽出

スナップショットBの視覚化結果を図表 22 に示す。この視覚化からは、4つの赤色の目立つ領域と、その他の小さな領域が数多く確認できる。4つの目立つ領域と、小さな領域から1つを選択し、スナップショット B から視覚的に抽出した標本（付録 図表 104 から図表 108 に掲載）とした。これらの標本を調べると、リンクコレクション（図表 104, 図表 105, 図表 106）、検索エンジン（図表 107）、およびホスティングサービス（図表 108）であった。スナップショット B による視覚化では、この期間においてダークウェブで支配的であるノードが4つ存在する様子が示されている。

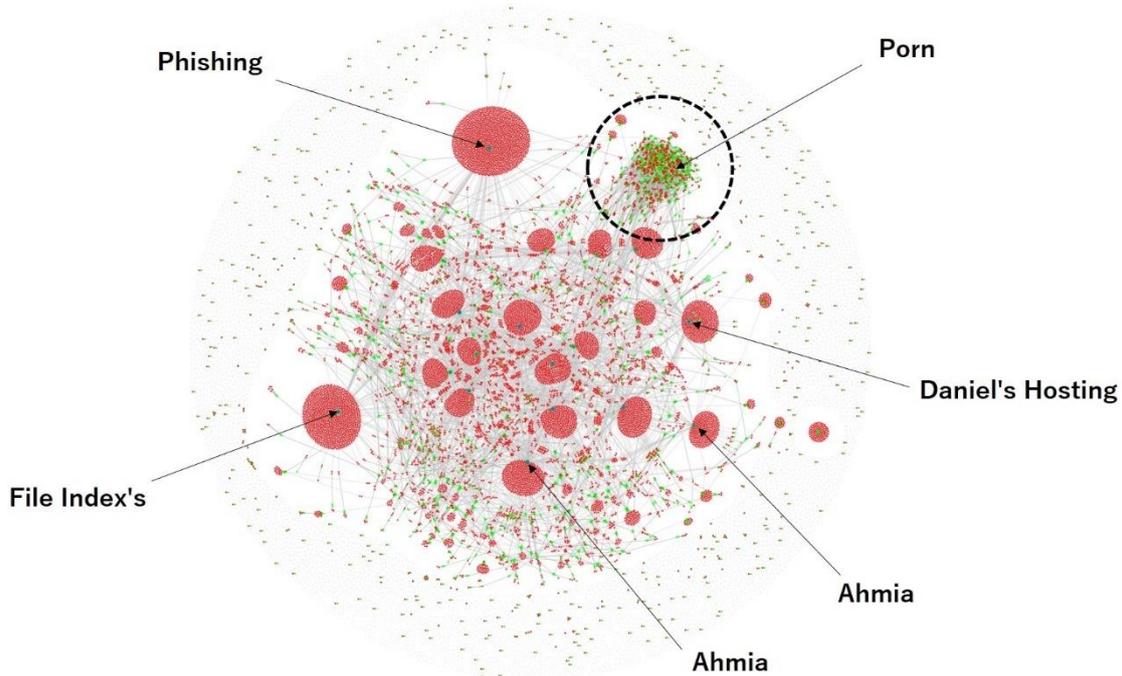


図表 22. スナップショット B のハイパーリンクネットワークグラフの視覚化

Ahmia, Fresh Onions, File Index's, Daniel's Hosting はそれぞれ固有名詞

第3項 スナップショットCの視覚的標本の抽出

スナップショットCの視覚化結果を図表 23 に示す。この視覚化からは、2つの赤色の目立つ領域、1つの緑色の目立つ領域、その他多数の小さな赤色の領域が確認できる。これらの目立つ領域のいくつかを、スナップショット C から視覚的に抽出した標本（付録 図表 109 から図表 114 に掲載）とした。これらの標本を調べると、リンクコレクション（図表 114）、検索エンジン（図表 111, 図表 112）、ホスティングサービス（図表 110）、ポルノ（図表 109）であった。スナップショット C の視覚化では、以前のスナップショットでは観察されなかった大きな緑色の塊（図表 109）が観察された。この領域を調べたところ、ポルノコンテンツを提供するノードが多数含まれていることがわかった。視覚的に抽出した標本の 1 つ（図表 113）は、この期間にダークウェブで人気のあったリンクコレクションサイト（図表 114）を模倣したフィッシングサイトであった。

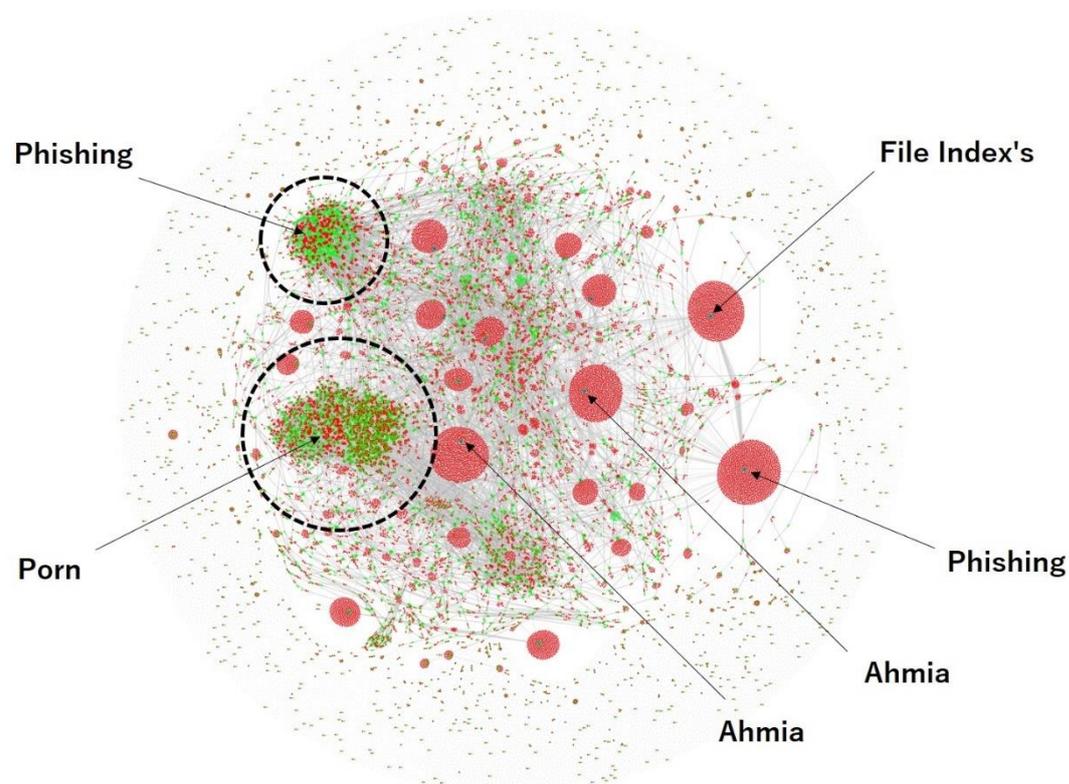


図表 23. スナップショット C のハイパーリンクネットワークグラフの視覚化

Ahmia, File Index's, Daniel's Hosting はそれぞれ固有名詞。 Porn, Phishing は一般名詞。

第4項 スナップショットDの視覚的標本の抽出

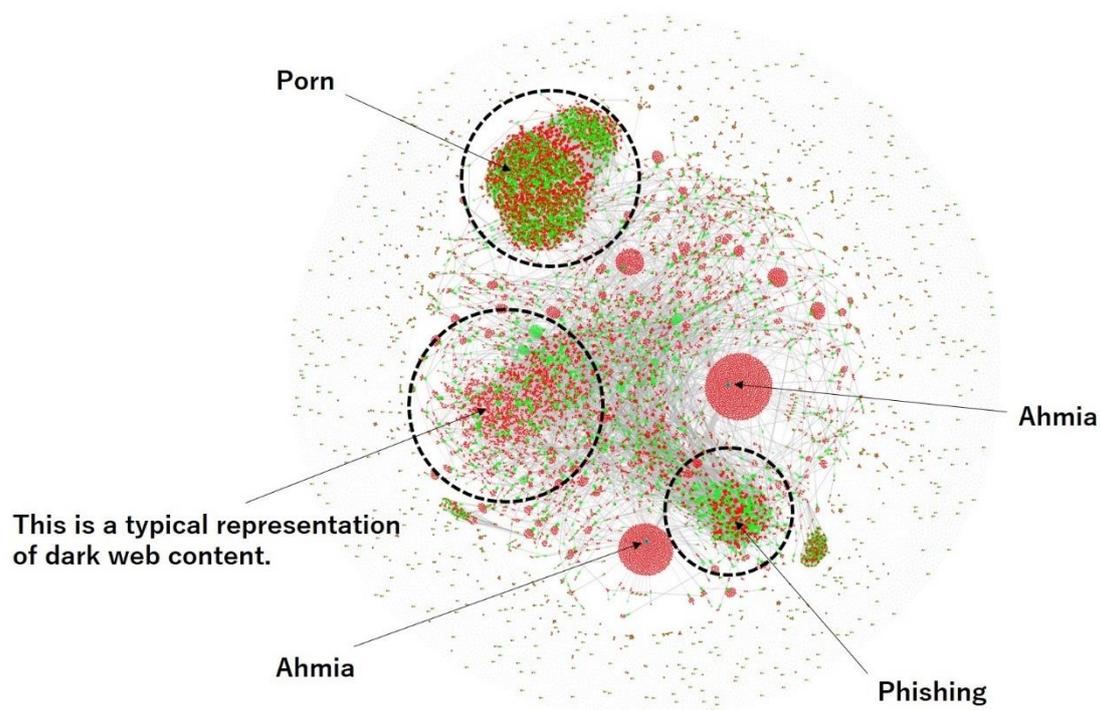
スナップショットDの視覚化結果を図表 24 に示す。この視覚化からは、4つの赤色の目立つ領域、2つの緑色の目立つ領域、いくつかの小さな赤色の領域が確認できる。これらの目立つ領域のいくつかを、スナップショット D から視覚的に抽出した標本（付録 図表 115 から図表 120 に掲載）とした。これらの標本を調べると、リンクコレクション（図表 115）、検索エンジン（図表 117、図表 118）、ポルノ（図表 120）、フィッシング（図表 116、図表 119）であった。スナップショット C の視覚化と比較して、ポルノに対応する領域（図表 120）はスナップショット D でさらに大きな範囲を占める。スナップショット D では、フィッシングに関連する領域（図表 119）を新たに確認した。このフィッシングの構造は、リンクコレクション（図表 115）を模倣する図表 116 の標本とは構造が異なり、マーケットを模倣する多数のドメインを生成し、お互いを少ないリンク数で相互にリンクさせていた。スナップショット ABCD を比較すると、スナップショット D は他と異なり、全体的に緑色が強く見える。この視覚化において、緑の濃度が高い場合、ハイパーリンクソースを持つノードが密集していることを意味している。つまり、この結果は、ネットワーク内のノードがハブノードと一方的にリンクされている傾向を持つスナップショット ABC とは異なり、スナップショット D では、より多くのエンドノードが相互にリンクされている様子を視覚的に示している。また、スナップショット D では、スナップショット ABC と同様に多くのハイパーリンクを持つノードの特徴である赤い塊もまばらに視認できる。



図表 24. スナップショット D のハイパーリンクネットワークグラフの視覚化
Ahmia, File Index's はそれぞれ固有名詞. Porn, Phishing は一般名詞.

第5項 スナップショットEの視覚的標本の抽出

スナップショットEの視覚化結果を図表 25 に示す。この視覚化からは、2つの赤色の目立つ領域、2つの緑色の目立つ領域、そして赤と緑が混合した境界のあいまいな広い領域が確認できる。これらの目立つ領域のいくつかを、スナップショットEから視覚的に抽出した標本（付録 図表 121 から図表 125 に掲載）とした。これらの標本を調べると、検索エンジン（図表 121, 図表 123）、ポルノ（図表 124）、フィッシング（図表 122）であった。スナップショットEの視覚的な結果は、スナップショットCとDの間で起こった変化の方向性を明確に特徴づけた。スナップショットCDの視覚化と比較して、ポルノとフィッシングの領域（図表 124, 図表 122）はスナップショットEでさらに目立っており、赤い領域はネットワーク全体でさらに目立たなくなった。つまり、リンクを多く持つハブノードの数はネットワークから全体的に減少している傾向が確認できた。スナップショットCDと比較して、スナップショットEでは相互にリンクされた状態を表す赤と緑の混合領域（図表 125）が更に拡大している。これらの混合領域は、次数のそれほど高くない特徴の薄いノードの集まる領域であるとして視覚的に認識できる状態になっている。混合エリア内のいくつかのノードをサンプリングして調べたところ、標準的なダークウェブコンテンツであることが確認できた。スナップショットEの視覚化からは、ダークウェブの多くのノードがハブノードに依存せず、相互に弱く結びつきはじめた様子を示している。



図表 25. スナップショット E のハイパーリンクネットワークグラフの視覚化
Ahmia は固有名詞. Porn, Phishing は一般名詞.

第5節 グラフによる視覚化の評価

ここでは、視覚化した結果から抽出した標本について評価するため、視覚的に抽出した標本の特徴評価とグラフによる視覚化結果と中心性指標との比較について確認する。比較のため、スナップショットからグラフ中心性指標の導出と上位10位のハブノードの抽出を行った。

第1項 視覚的に抽出した標本の特徴評価

視覚的に抽出した結果の妥当性について評価を行う。評価方法はTakahashiらによる色彩特徴の分類・類型化手法 [46]を参考にした。Takahashiらは実世界における色彩分類の指標として中間色を軸とした色空間を使用したが、本評価では、主要な色彩がRG成分とその混合色であることがあらかじめ明らかである本研究の視覚化特性を加味して、RGBを軸とする色空間により定量化した上で評価した。評価には、期間中最も複雑な形状を記録したスナップショットDから視覚的に抽出した大きさの異なる21種類の標本(付録C1図表75から図表95に掲載)を使用した。これを基準標本群とする。

基準標本群における各標本にA~Uまでのラベルをつけ、それぞれの標本に対して背景領域を除いて正規化したRGB成分構成比率を求め定量化した(図表26)。典型的な領域はラベルA・B・Cである。ラベルD・Eも領域と認識できる。領域ではない標本はラベルO・P・Qである。ラベルR・Sは、密集はしているが、色成分としての境界はあいまいである。

次に、基準標本群における各標本に対して背景を除く色成分を抽出し、赤色成分、緑色成分、青色成分それぞれの占める比率を求める。RGB各成分の比率を三角グラフにプロットすると、おおよそ4つの性質の異なる群が明らかとなった(図表27)。A群はサンプリング範囲を広くとった結果として視覚的にも特徴が判然としない標本群である。ラベルR・SはA群に属する。B群は赤色の特徴が強く出る領域をサンプリングした群である。ラベルA・B・CはB群に属する。C群は緑色の特徴が強く出る領域をサンプリングした群である。ラベルD・EはC群に属する。D群はサンプリング範囲を小さくとった結果として個々の関係性にフォーカスされた標本群である。ラベルO・QはD群に属する。ラベルPは各成分値が0であるため、三角グラフにはプロットされない。B群及びC群に属する場合、色成分の点から特徴があるといえる。D群は領域ではない疎な領域であるという特徴を示す。A群に属する場合、色成分としての特徴は無いといえる。

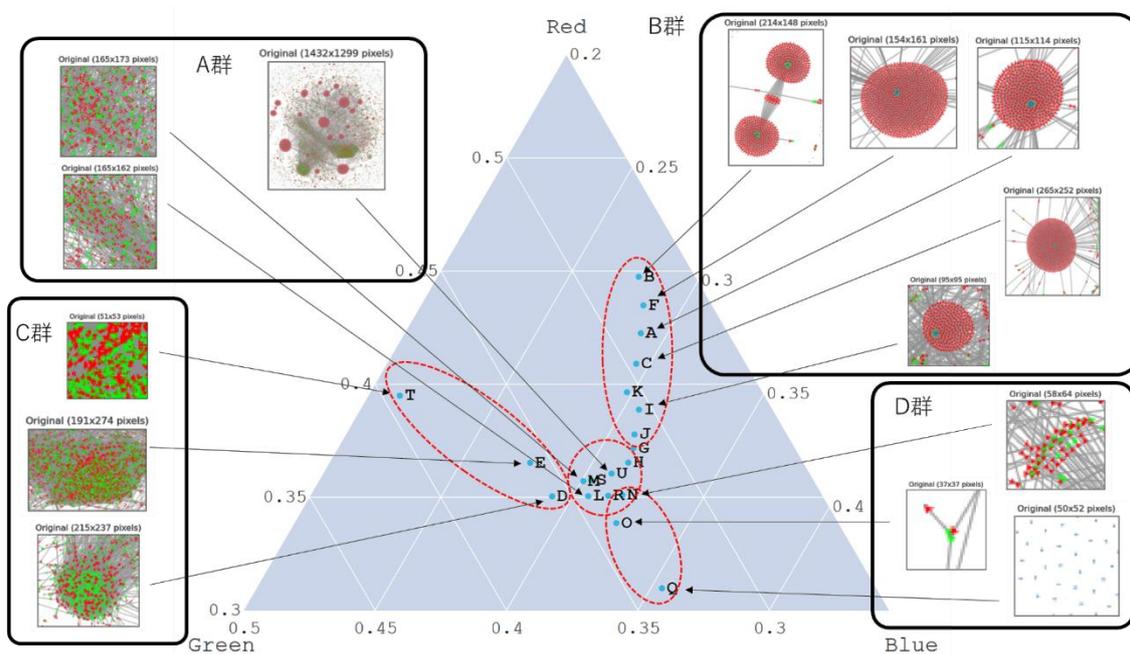
人間は目視によって色成分の情報だけではなく輪郭情報を認識することが出来る。目視により目立つ領域として抽出できる場合、機械的に領域を抽出することを目的とする領域

判定アルゴリズムでも同様に抽出できる可能性が考えられる。これを確認するため、基準標本群の各標本は、Watershed 法 [47]により領域を自動判定し、境界情報を標本画像に上書きした（付録 C1 図表 75 から図表 95）。Watershed による領域の計算には OpenCV [48]の実装を使用した。Watershed 法そのものの評価については Kornilov らの研究 [49]を、領域分割の原理については CMM の解説 [50]をそれぞれ参照されたい。

図表 26. 視覚的に抽出した 21 標本の RGB 成分構成比率

背景領域を除いて正規化

Label	Number of pixels	Red rate	Green rate	Blue rate
A	13,110	0.423	0.288	0.290
B	31,672	0.448	0.276	0.277
C	66,780	0.409	0.296	0.295
D	50,955	0.350	0.357	0.292
E	52,334	0.365	0.358	0.276
F	24,794	0.435	0.281	0.285
G	37,810	0.372	0.316	0.312
H	140,220	0.365	0.321	0.314
I	7,920	0.389	0.305	0.306
J	9,025	0.378	0.312	0.310
K	2,760	0.396	0.306	0.298
L	26,730	0.351	0.344	0.306
M	28,545	0.357	0.342	0.301
N	3,712	0.351	0.330	0.319
O	1,369	0.339	0.339	0.323
P	6,084	n/a	n/a	n/a
Q	2,600	0.310	0.336	0.354
R	95,976	0.351	0.336	0.313
S	125,132	0.358	0.339	0.303
T	2,703	0.395	0.393	0.212
U	1,860,168	0.360	0.330	0.310



図表 27. 基準標本群の分布と分類



図表 28. 基準標本群の各標本に占める RGB 成分の比率

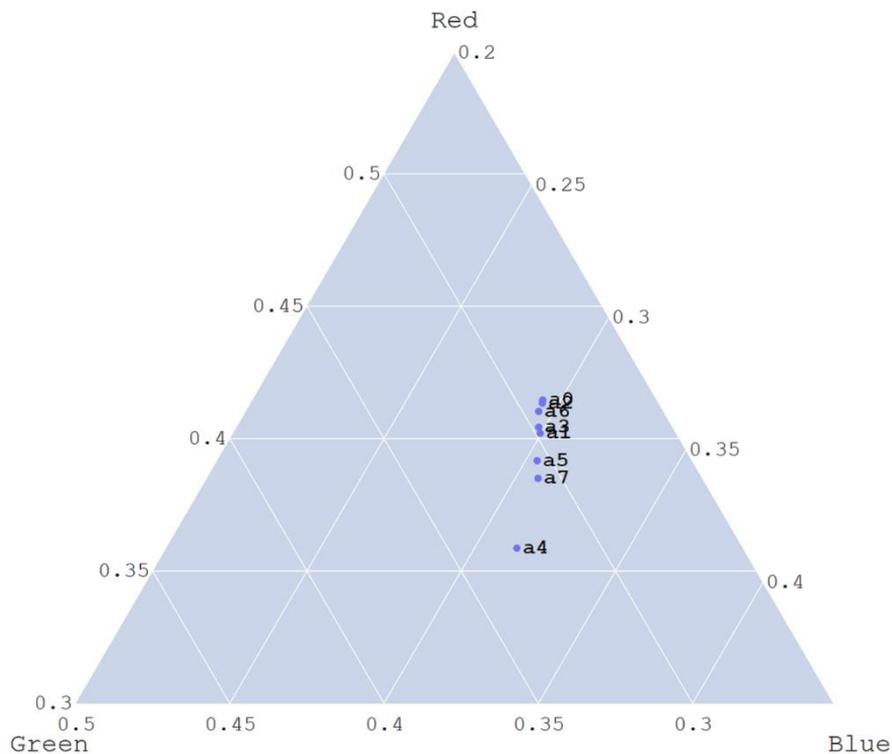
背景領域を除いて正規化。RGB の順で降順に並べ替えている。

視覚的に抽出として最も典型的な分類は B 群である。赤い塊が形成されるためグラフ上で目立つ。同様に、緑の塊が形成される C 群も同様に目立つ。視覚的な抽出から領域として選択された領域を示すラベル D (緑色成分比率 0.357)・E (緑色成分比率 0.358)・T (緑色成分比率 0.393) についてはいずれも C 群に属し、緑色成分の比率が 0.35 以上の値を示しており、他の標本と比べて高い値を示している。その他に分布上で特徴を持つ D 群は、周囲にノードが集まらない疎な特徴を持つ。視覚的には白の成分が目立つ。A 群に属する U・M・L については色が拮抗し合い、視覚的にも境界があいまいとなり目立つとはいえない。図表 28 からも、各 RGB 値が近い値を取り、何かの色成分が突出する状態ではない。

スナップショット A から視覚的に抽出した各領域（付録 図表 96 から図表 103 に掲載）について同様に RGB 成分を求める（図表 29）。三角グラフにプロットすると、標本は B 群に属するものが多くみられた（図表 30）。

図表 29. スナップショット A から視覚的に抽出した各標本の RGB 成分

Label	Number of pixels	Red rate	Green rate	Blue rate
a0	30100	0.415	0.291	0.294
a1	27874	0.402	0.298	0.300
a2	40425	0.413	0.292	0.295
a3	27221	0.404	0.298	0.298
a4	17920	0.359	0.328	0.314
a5	22032	0.392	0.305	0.304
a6	31832	0.410	0.295	0.295
a7	21895	0.385	0.308	0.307

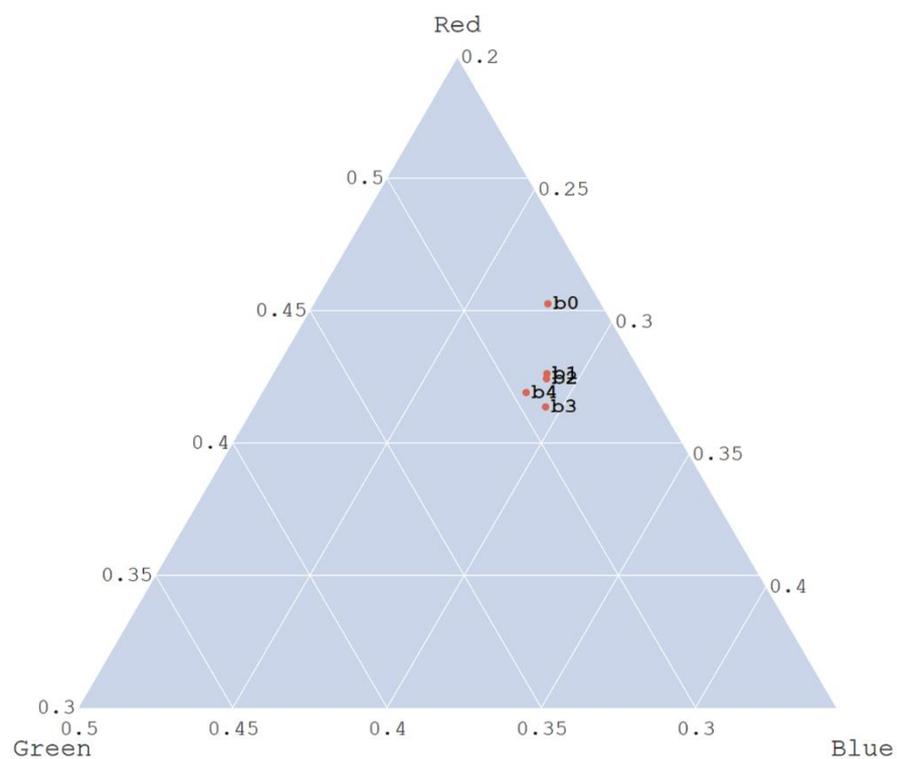


図表 30. スナップショット A から視覚的に抽出した各標本の分布

スナップショット B から視覚的に抽出した各領域(付録 図表 104 から図表 108 に掲載)について同様に RGB 成分を求める (図表 31). 三角グラフにプロットすると, 標本は B 群に属する傾向が強く表れた (図表 32).

図表 31. スナップショット B から視覚的に抽出した各標本の RGB 成分

Label	Number of pixels	Red rate	Green rate	Blue rate
b0	43800	0.453	0.272	0.276
b1	36465	0.426	0.285	0.289
b2	43056	0.424	0.286	0.290
b3	36472	0.414	0.292	0.295
b4	12862	0.419	0.295	0.286

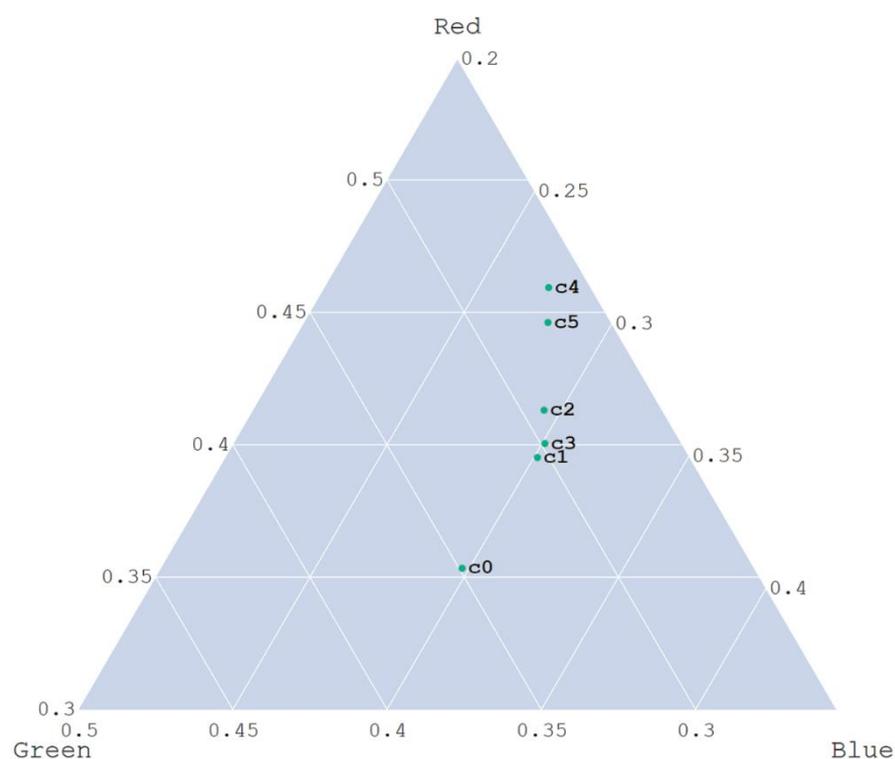


図表 32. スナップショット B から視覚的に抽出した各標本の分布

スナップショット C から視覚的に抽出した各領域(付録 図表 109 から図表 114 に掲載)について同様に RGB 成分を求める (図表 33). 三角グラフにプロットすると, 標本は B 群に属するものが多くみられたが, c0 については成分の傾向が異なり, A 群に近い場所にプロットされた (図表 34).

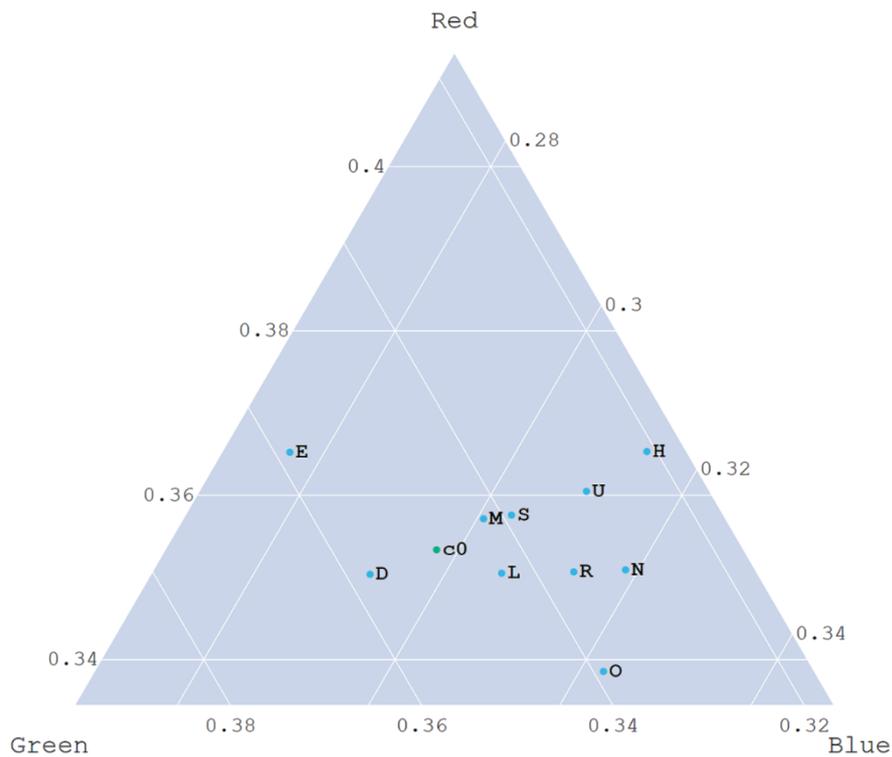
図表 33. スナップショット C から視覚的に抽出した各標本の RGB 成分

Label	Number of pixels	Red rate	Green rate	Blue rate
c0	71554	0.353	0.349	0.298
c1	39402	0.395	0.304	0.301
c2	21009	0.413	0.293	0.294
c3	23023	0.400	0.299	0.301
c4	66250	0.459	0.268	0.273
c5	46620	0.446	0.275	0.279



図表 34. スナップショット C から視覚的に抽出した各標本の分布

ラベル c0 について RGB 成分に分解した様子を図表 109 に示す。三角グラフに基準標本群の分布とスナップショット C の分布をプロットし、ラベル c0 の周辺にフォーカスすると、ラベル M・L・D の中心に位置することがわかる（図表 35）。これは A 群または C 群のおおよそ中間の値を取るため、色成分の点からは区別は困難であった。しかしながら、Watershed 法により自動的に判定した結果は、目立つ領域が認識された。

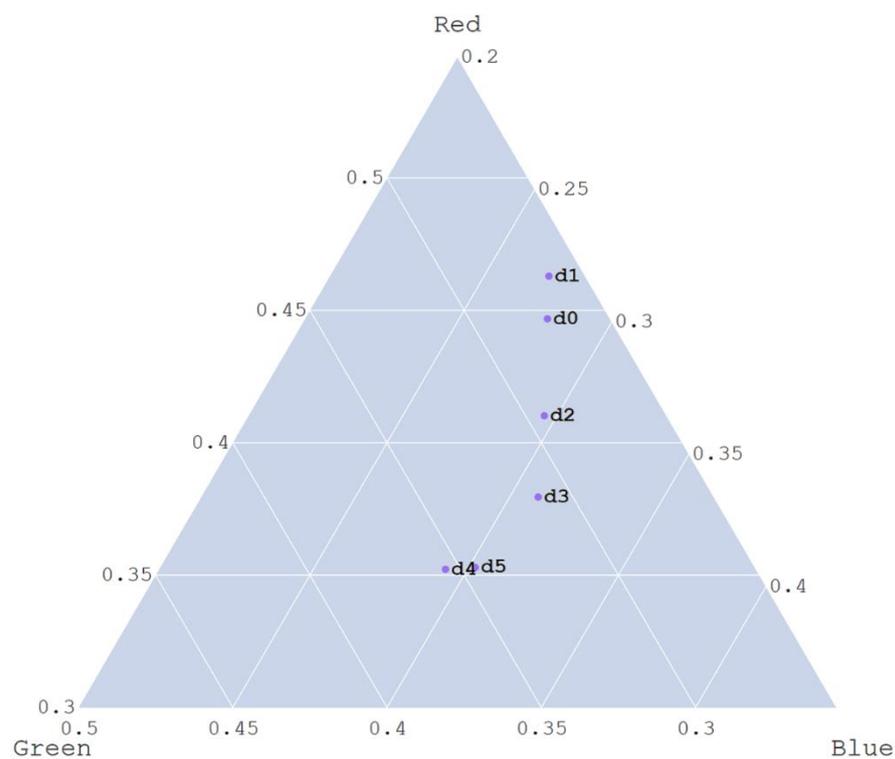


図表 35. ラベル c0 周辺の標本の分布

スナップショット D から視覚的に抽出した各領域(付録 図表 115 から図表 120 に掲載)について同様に RGB 成分を求める (図表 36). 三角グラフにプロットすると, 標本は B 群に属するものが多くみられたが, d4・d5 については A 群に近い場所にプロットされた(図表 37).

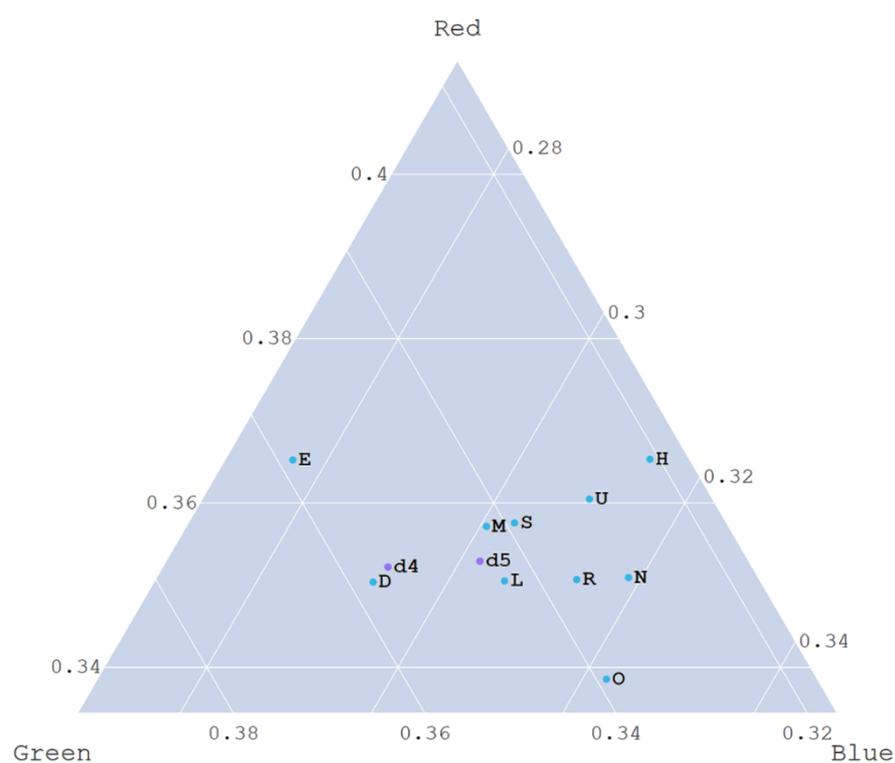
図表 36. スナップショット D から視覚的に抽出した各標本の RGB 成分

Label	Number of pixels	Red rate	Green rate	Blue rate
d0	33670	0.447	0.275	0.279
d1	36848	0.463	0.266	0.271
d2	31476	0.410	0.294	0.296
d3	52460	0.379	0.311	0.309
d4	72800	0.352	0.355	0.293
d5	141984	0.353	0.345	0.302



図表 37. スナップショット D から視覚的に抽出した各標本の分布

ラベル d4・d5 について RGB 成分に分解した様子を図表 119 及び図表 120 に示す。三角グラフに基準標本群の分布とスナップショット D の分布をプロットし、ラベル d4・d5 の周辺にフォーカスすると、ラベル M・L・D の間に位置することがわかる (図表 38)。ラベル d4 は C 群であるラベル D に近く、ラベル d5 については A 群であるラベル M・L に近い。Watershed 法により自動的に判定した結果は、ラベル d4 においては小さくまばらな領域を認識し、d5 についてはいくつかの密集した領域の存在を示した (図表 119, 図表 120)。

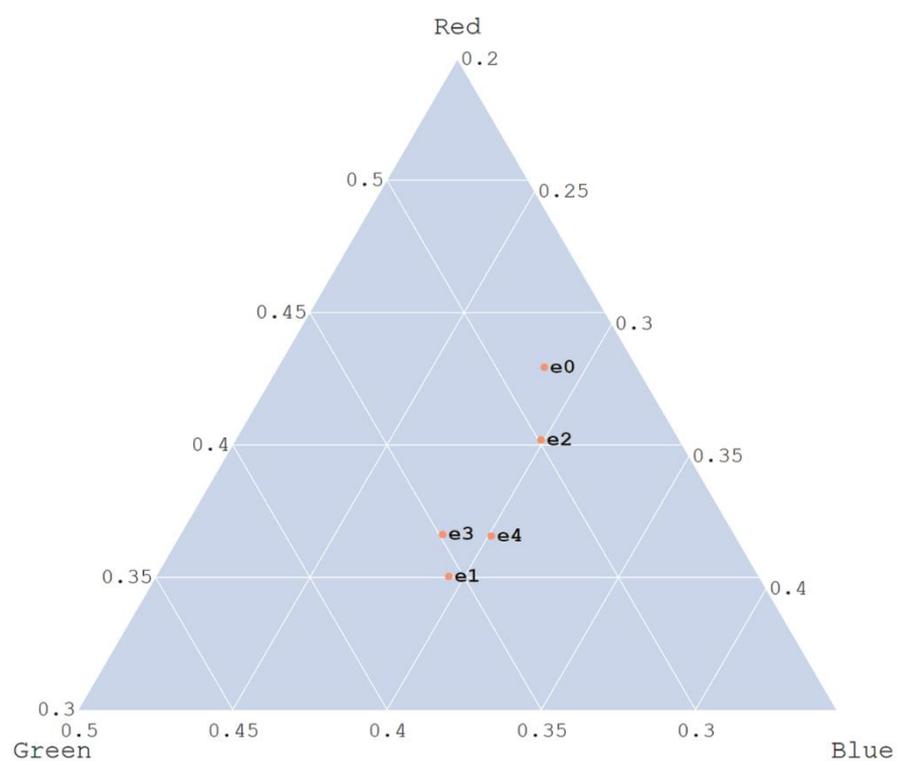


図表 38. ラベル d4・d5 周辺の標本の分布

スナップショット E から視覚的に抽出した各領域（付録 図表 121 から図表 125 に掲載）について同様に RGB 成分を求める（図表 39）。三角グラフにプロットすると、 $e0 \cdot e2$ は B 群の傾向を強く示し、 $e1 \cdot e3 \cdot e4$ は A 群に近い場所にプロットされた（図表 40）。

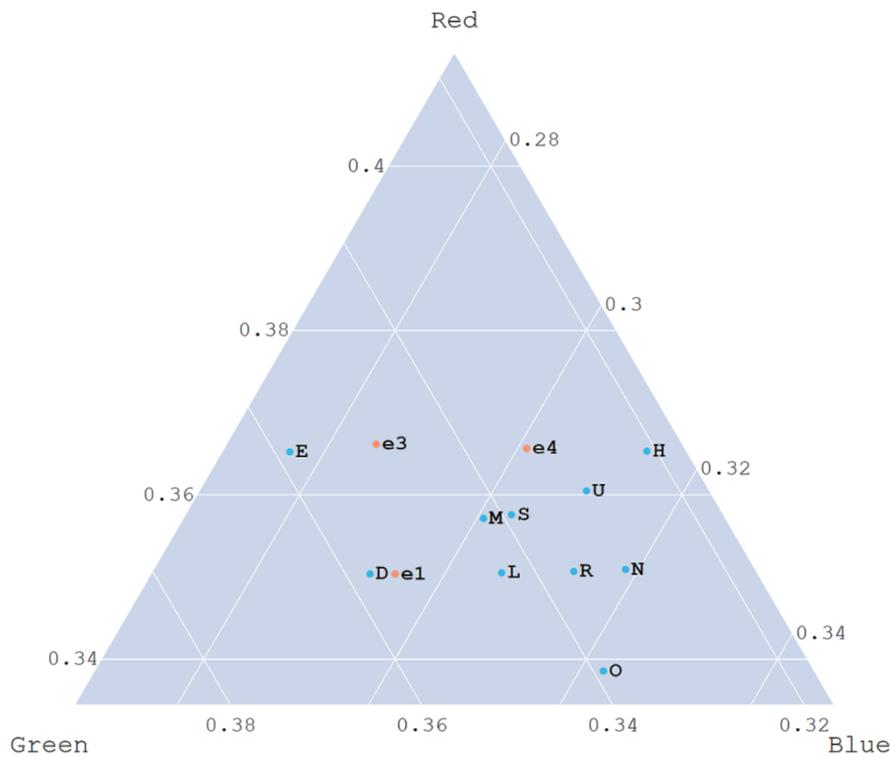
図表 39. スナップショット E から視覚的に抽出した各標本の RGB 成分

Label	Number of pixels	Red rate	Green rate	Blue rate
e0	45136	0.429	0.284	0.286
e1	89951	0.350	0.355	0.295
e2	52192	0.402	0.299	0.299
e3	207711	0.366	0.349	0.285
e4	186082	0.366	0.333	0.301



図表 40. スナップショット E から視覚的に抽出した各標本の分布

ラベル e1・e3・e4 について RGB 成分に分解した様子を付録の図表 122・図表 124・図表 125 にそれぞれ示す。三角グラフに基準標本群の分布とスナップショット E の分布をプロットし、ラベル e1・e3・e4 の周辺にフォーカスすると、ラベル D・U の間に位置することがわかる (図表 41)。ラベル e1 は C 群のラベル D に近く、ラベル e4 は A 群のラベル S・U の中間に位置する。ラベル e3 についてはこれらの分布からは少し離れた位置であった。Watershed 法により自動的に判定した結果は、ラベル e1・e3 についてははっきりとした領域を認識したものの (図表 122, 図表 124), e4 については小さくまばらな領域を認識した (図表 125)。



図表 41. ラベル e1・e3・e4 周辺の標本の分布

最後に、各スナップショットから目視で抽出した各標本について RGB 成分と Watershed 法による領域判定結果から、目視で抽出した標本の特徴の有無について比較を行った。比較の結果、目視抽出の結果では色成分としてはあいまいとされる A 群に属する標本 (c0, d5, e1, e3) においても、アルゴリズムによる判定では領域があるとして認識されることが確認できた。一方で、同じ A 群に属する標本でも、アルゴリズムによる判定では領域があるとは認識されづらいものが目視抽出により選ばれていることも確認した (d4, e4)。

これらの評価を行った結果、色情報と境界情報のいずれかで認識できる領域について、視

覚的な特徴があるとしてサンプリングしている傾向が示された。今回の視覚化においては、塊を認識すること、色情報を背景色である白と、各ノードの接続状態を示す赤と緑、そしてそれらの要素の疎密により表現されることから、これらのサンプリング結果の傾向は、視覚的に特徴的な標本として妥当であると判断できる。

第2項 グラフ中心性指標の導出とハブノードの抽出

グラフにおける重要なノードを調べるため、各グラフの中心性指標 [51] を計算した。中心性指標とは、ネットワークにおける重要なノードを計算により識別するために使用される値であり、次数中心性は、ノードに接続されているエッジの数で表される。有向グラフにおけるエッジには方向性があるため、次数中心性は入力方向と出力方向で別々に計算できる。入力次数はノードに向けられた内向きの辺の数であり、出力次数は他のノードに向けられた外向きの辺の数である。単に次数と表現する場合、入力と出力のそれぞれの辺の数合計である。ネットワーク分析のフレームワークとして人気のある Network X [52] を使用して、次数中心性 [51]、入力次数中心性 [51]、出力次数中心性 [51]、そして PageRank [53] をそれぞれ計算した。グラフ $G := (V, E)$ におけるノード数 n のグラフの頂点 v における次数中心性 C_d は以下の式で求めることができる。

$$C_d(v) = \frac{\text{deg}(v)}{n-1} \quad (2)$$

求めた中心性の指標ごとに、各スナップショットからスコアが最も高い上位 10 個のドメインを抽出し、各グラフの各中心性指標の上位 10 件のハブノードとした。中心性指標の上位 10 件の結果は、付録に掲載している（付録 図表 54 から図表 74）。

第3項 グラフによる視覚化結果と中心性指標との比較

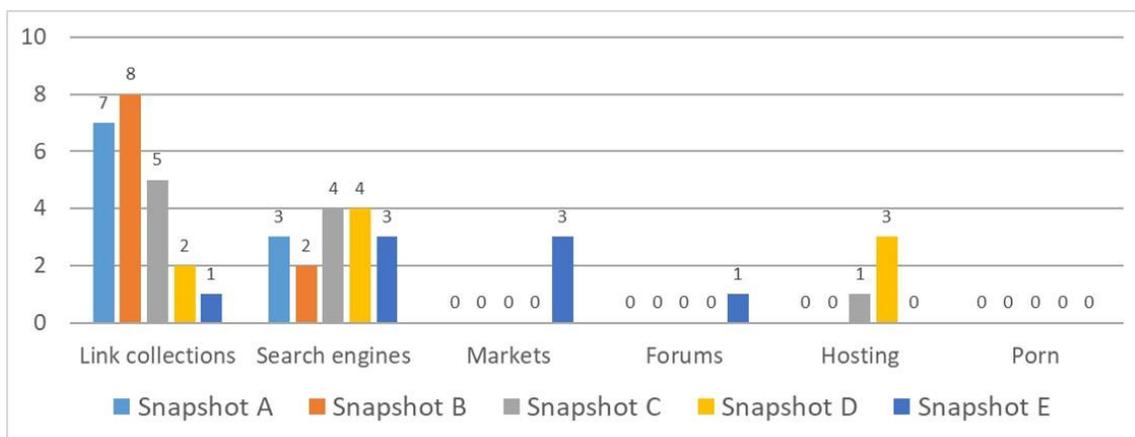
グラフによるネットワーク構造から目視により抽出した結果を、中心性指標と比較した。

スナップショット A から計算された次数中心性を使用し、上位 10 ノードのカテゴリを特定した (図表 42)。識別されたカテゴリはリンクコレクションと検索エンジンであり、ホスティングを除いて、視覚的な結果はこれと一致した。さらに、出力次数と入力次数の両方の中心性について、スナップショット A の上位 10 ノードのカテゴリを特定した (図表 43 および図表 44)。出力次数と入力次数の中心性のカテゴリは、それぞれリンクコレクションと検索エンジン、およびリンクコレクション、マーケット、フォーラムであった。視覚的な結果は、ホスティング、マーケット、およびフォーラムのカテゴリを除いて一致した。同様に、PageRank を使用して上位 10 ノードのカテゴリ傾向を調査した (図表 45)。次数の中心性と同様に、PageRank のカテゴリ別の傾向はリンクコレクションと検索エンジンであった。

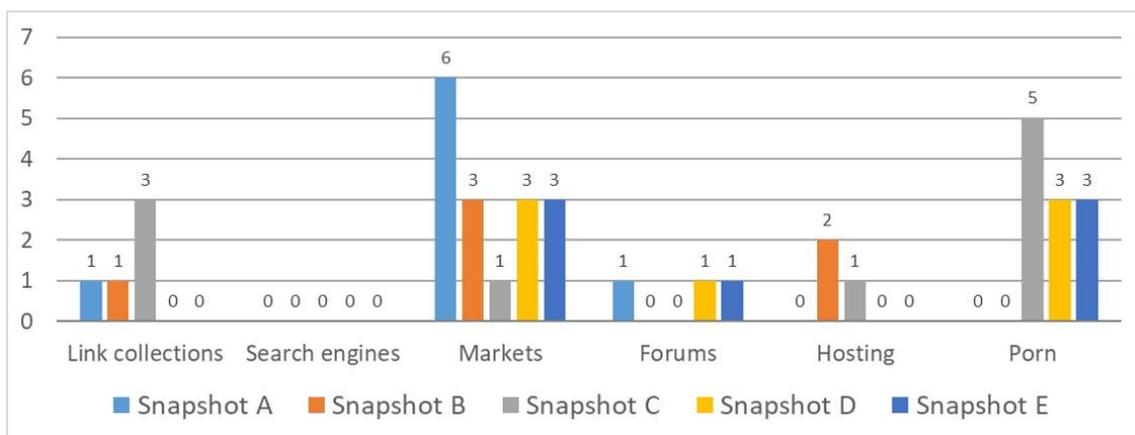
同様に、スナップショット B の各中心性指標からカテゴリを抽出すると、次数中心性と入力次数中心性から取得されたカテゴリはリンクコレクションと検索エンジン (図表 42 と図表 43) であり、PageRank から取得されたカテゴリはリンクコレクション、検索エンジン、およびホスティングであった (図表 45)。これらはすべて視覚的な結果と一致した。出力次数中心性から得られたカテゴリは、リンクコレクション、マーケット、およびホスティング (図表 43) であり、視覚的な結果はマーケットを除いて一致していた。

スナップショット C の各中心性指標からカテゴリを抽出したところ、次数中心性、入力次数中心性、および PageRank から得られたカテゴリの傾向は、リンクコレクション、検索エンジン、およびホスティングであった。ポルノを除いて、視覚的な結果は一致した。一方で、出力次数中心性から得られたカテゴリの傾向は、リンクコレクション、マーケット、ホスティング、ポルノであり、視覚的な結果は検索エンジンとマーケットカテゴリだけが一致していた (図表 43)。

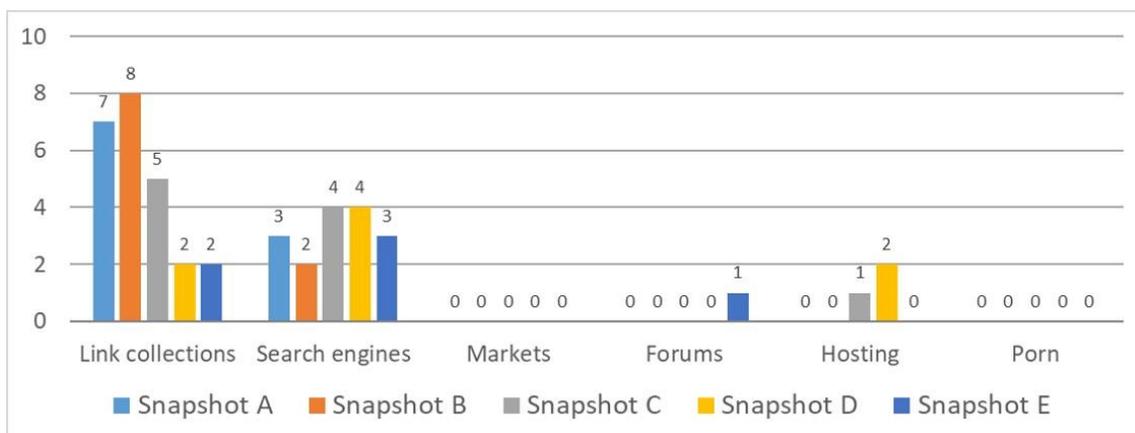
スナップショット D の各中心性指標からカテゴリを抽出した。次数中心性、入力次数中心性、および PageRank から得られたカテゴリの傾向は、リンクコレクション、検索エンジン、およびホスティングであり、視覚的な結果は、ホスティングとポルノのカテゴリを除いて一致した (図表 42, 図表 44, および図表 45)。出力次数中心性から得られたカテゴリの傾向は、マーケット、フォーラム、ポルノであった。ポルノカテゴリのみが視覚的な結果との一致を示した (図表 43)。



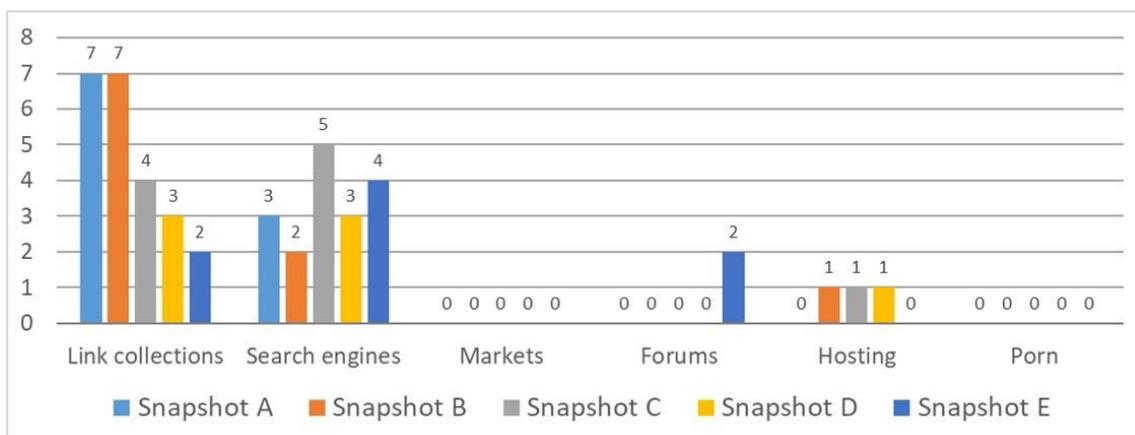
図表 42. 次数中心性の上位10ドメインの占めるカテゴリの集計



図表 43. 出力次数中心性の上位10ドメインの占めるカテゴリの集計



図表 44. 入力回数中心性の上位10ドメインの占めるカテゴリの集計



図表 45. ページランクの上位10ドメインの占めるカテゴリの集計

スナップショット E の各中心性指標からカテゴリを抽出した。次数中心性から得られたカテゴリの傾向は、リンクコレクション、検索エンジン、マーケット、およびフォーラムであり、検索エンジンのカテゴリを除いて、視覚的な結果は一致しなかった (図表 42)。入力次数中心性と PageRank から得られたカテゴリの傾向は、リンクコレクション、検索エンジン、フォーラムであり、検索エンジンのカテゴリを除いて、視覚的な結果は一致しなかった (図表 43 と図表 45)。出力次数中心性から得られたカテゴリの傾向は、マーケット、フォーラム、ポルノであり、視覚的な結果はポルノのカテゴリにのみ一致した (図表 44)。

これらの比較により、いくつかの中心性指標の結果は、視覚的表現の結果と一致していることが分かった。視覚化による抽出では、中心性指標だけでは認識が困難な領域を容易に認識することができた。つまり、ネットワーク全体で発生した大規模な構造の変化を視覚化では容易に捉えたといえる。

各ネットワーク中心性指標の結果は、上位 10 のハブノードのリンクコレクションの数が時間の経過とともに減少することを示した。一方、上位 10 のハブノードの検索エンジンの数は、測定期間中に大幅に変化したようには見えない。各中心性指標から得られた結果は、ネットワーク内のリンクコレクションの影響の減少と比較して、相対的に検索エンジンの影響が増加したことを示した。これは、視覚化により把握できた傾向と一致した。

第6節 議論

視覚化によって確認された観察期間中（2018年6月1日から2021年1月30日まで）の主な変化は、ダークウェブ上のリンクコレクションの重要性が低くなり、相対的に検索エンジンの重要性が高くなったことである。

リンクコレクションには通常多くのリンクがあるため、ハイパーリンク関連のグラフをプロットすると、それらは出力次数の高いノードとして表示される。本研究の視覚化では、リンクコレクションは大きくて目立つ赤い領域として表示されている。最初のスナップショットから得られた視覚化では、多くの赤い領域が表示されていた（図表 21）。この結果は、ダークウェブ上の多くのノードにリンクがないことを述べた Griffith らによって 2017 年に報告された分析結果と同様であった [23]。

スナップショットが進むにつれ、ダークウェブでは大きな赤い領域を持ついくつかのノードが支配的になった（図表 22）。ポルノコンテンツの大きな領域が 2019 年 5 月にダークウェブに登場した（図表 23）。この頃、ダークウェブのグラフ構造も変化を始めた。ポルノコンテンツの領域は相互に弱く結びつき合い、リンクコレクションほど大きな出力次数は持っていない。中心性指標では目立たないこれらの関係性は、辺の方向を赤と緑の色で示すことによって視覚的に表されている。

ポルノコンテンツの領域は、観察期間の後半には拡大を続けた（図表 23, 図表 24 及び図表 25）。スナップショット D を視覚化した結果からいくつかの特徴的なノードを調べている間、この期間中にダークウェブ上に大規模なフィッシングが観察された（図表 25）。ポルノコンテンツの領域と同様に、この悪意のあるネットワークの構造には中央となるハブノードが無かった。言い換えれば、今回比較を行った 3 つのネットワーク中心性指標でそれらを見つけることは困難である。対照的に、本研究の視覚化では、これらの不正な領域を緑色で特徴づけられる領域として視覚的に認識することができる。

このような不正な領域の存在は、今日のダークウェブにおいて悪い影響を与える特徴の一つであり、すなわち悪意のあるノードがダークウェブの持つ高い匿名性を悪用して、排除されることなくダークウェブに長期間とどまることができることを意味している。この結果は、現在の匿名通信システムの脆弱性を浮き彫りにしているともいえ、匿名通信システムをプライバシー保護という匿名性の観点から改善する必要があるのであれば、これらの悪意あるネットワークを排除してダークウェブの匿名性を保護できる防御メカニズムが将来的に必要であることを示唆している。

ここからはいくつかの関連研究との差異について議論する。Brunner ら [22] は、収集したデータの 10% にあたるおおよそ 1,000 ノードのリンク関係をもとにダークウェブのグラフを作り視覚化を行った。彼らの視覚化では、グラフの頂点と辺の色を PageRank の結果を元に緑から紫のグラデーションによって色付けを行い、PageRank の高いノードが紫で表示

されるように工夫している。彼らの視覚化について詳しい説明は無いため、グラフィレイアウト方式については明らかではない。本研究による視覚化は48,360(スナップショット A)～103,674 (スナップショット E) 及び 172,740 (スナップショット A-F) ノードの視覚化を行っているため、同時に視覚化されるノード数がおおよそ 48 倍～172 倍と大きく異なる。また、彼らは時系列の視覚化は行わなかった。

Hyperion Grey [29]は、2019年に3,700のTorオニオンサービスの視覚化を作成した。本研究で扱う視覚化と比べるとノード数が極めて小さい。彼らの視覚化には、ダークウェブサイトのスクリーンキャプチャをグラフ構造にマッピングする手法が含まれていた。本研究の視覚化にはスクリーンキャプチャは含まれていない点が彼らの視覚化と異なる。彼らの視覚化はサイト構造の類似性をもとに関係グラフ構造を定義しているため、ハイパーリンク関連のグラフではなかった。本研究はハイパーリンク関係のグラフ構造を視覚化している点が彼らの視覚化とは異なる。

Cilleruelo ら [28]は様々なデータセットを混合してダークウェブのリンク関係を視覚化した。彼らは視覚化によって、特定のノードを解析することは困難であるとしながらも、I2P ネットワークには明確なサブネットワークを形成するクラスタが存在することを発見した。彼らの視覚化には、グラフィレイアウトに関する詳細な説明は無いものの、Tor と I2P の両方を含む 49,249 ノードが視覚化され、それぞれ明るい灰色 (Tor ノード) と濃い灰色 (I2P ノード) で色分けの工夫がなされている。視覚化されたデータ数としては我々の最も小さいデータである 48,360 (スナップショット A) の視覚化と同程度の規模が視覚化されているものの、172,740 (スナップショット A-F) ノードの視覚化と比べると小さい。本研究の視覚化は Tor のみである点も彼らの視覚化と異なる。また、彼らは時系列の視覚化は行わなかった。

いくつかの関連研究と本研究との比較を行った結果、ダークウェブのグラフ構造の視覚化は積極的に行われているとはいえ、時系列で視覚化しているものもみられなかった。大規模なグラフ構造の視覚化には高い解像度が必要となる点や、グラフ描画における計算コストが高い点、視覚化による比較方法が本研究により示されるまで確立していなかった点がこれまで積極的に視覚化されてこなかった理由と考えられる。

第7節 本章のまとめ

本章では、第3章で作成したスナップショットの有向グラフデータについて、ばねモデルを用いて視覚化し、その結果を読み解く道筋を示した。視覚化において、エッジの始点と終点でそれぞれ異なる色をつけることで、本研究特有の視覚化効果を得た。この視覚効果からは、ダークウェブグラフにおけるハブノードの存在と、低い次数で多くのノードが結合する領域の存在を浮き彫りにする。更に、視覚化結果の評価のために、これらの視覚化結果を、ネットワークグラフの代表的な中心性指標と比較を行った。中心性指標は、グラフデータから計算により自動的に求めることが出来る。結果として、中心性指標からは特徴があるとはみなされないいくつかの領域を視覚化により発見することができた。更に、これらの視覚化結果を時系列で比較した結果、ダークウェブにおける情報検索の在り方が、リンクコレクションを中心としたハブ構造の関係性をもつ形から、検索エンジンを中心とする相互のリンクの関係性をもつ構造へと徐々に変化している様子が明らかになった。視覚化による領域からは、ダークウェブの中でも大きな勢力を持つポルノネットワークやフィッシングの存在が確認できた。分析の結果、これらの勢力は観測を開始した当初は存在せず、2019年5月以降に誕生したことが明らかとなった。

第5章

ダークウェブネットワークのノード数の推定

第1節 概要

本章では、ダークウェブネットワークのノード数の推定について記載する。ダークウェブの大きさを求めることは困難であると考えられており、我々の知る限り最大規模の観測を行った本研究においても、ダークウェブの全てを瞬時に観測することは実現できていない。そこで、本章ではダークウェブネットワークのノード数に言及のある関連研究を紹介したあと、観測データをもとに、Peterson による標識再捕獲法を用いてダークウェブの大きさの推定を試みた。Peterson による標識再捕獲法は個体群生態学の分野で野生動植物の個体数の推定を行う場合にしばしば使用される手法である。野生動植物の個体数も、あるフィールドにおいて全ての個体を同時に識別してカウントすることは困難であるため、ダークウェブの大きさの推定と同質の問題を抱えている。

第2節 関連研究

ここでは、ダークウェブのネットワークの大きさの推定を行うにあたり、ウェブ及びダークウェブの大きさについて言及のあるものを中心に関連研究を示す。

Coffman は [54], 1998 年にウェブの大きさと成長率に関する初期の分析を行った。彼らは、パブリックインターネットのトラフィックに焦点を当て、トラフィックの観点からウェブの大きさと成長率の測定を試みた。彼らはインターネットの規模と成長率について、包括的な情報源に乏しいことを指摘し、インターネットの大きさを測る研究の多くは、ユーザー数を対象としていることと、インターネットに接続されているコンピュータの数、すな

わちノード数を対象としていることに言及した。彼らは研究の中で、インターネットは緩い集合体であるため、サイズを見積もる際に何を含めるかを決定することは困難であると述べている。彼らは研究の中でインターネットのノード数に言及しているものの、既存研究を引用した測定値を示しただけであり、推定はしなかった。

Loesing ら [33]は、2010年にダークウェブの統計を分析した。彼らは研究の中で、ダークウェブを測定することは、ダークウェブの利用される国を把握することや匿名通信システムのパフォーマンスを改善させること、そして匿名通信システムの利用目的を調べることに役に立つと述べている。彼らはダークウェブの統計として、Torシステムに固有の測定情報として、Torネットワークに接続するクライアント国情報と、ポートやトラフィックに関する測定を行い、統計情報を示した。彼らの研究はTorシステムにより密接な部分に注力していたため、Torネットワークを構成するTorノードへの言及を行ったものの、Torネットワークを利用してその中に形作られるダークウェブやそのコンテンツの観点からの言及は行わなかった。

Brunner ら [22]は、2018年にダークウェブのネットワーク構造を詳細に分析した。彼らは67,296,302のリンクから34,714のTorドメインを見つけ、約10,000のアクティブなサイトをクロールして収集したデータを分析した。彼らは、クロールングにより発見した34,718のアドレスについて、Tor metricsにより推定された70,640と比較して約半数のダークウェブが明らかになったと仮定できると主張している。しかしながら、彼らの収集したアドレスが、その時点で有効なアドレスであったかどうかは判明せず、データが収集された時期についても判然としない。

Filleruelo ら [28]は、2020年にTorとI2Pネットワーク間の接続を明らかにするために、ダークウェブネットワークのグラフ構造を作成した。彼らは、Ahmia [32]およびDUTA-10Kデータセット [12]、ダークウェブからの独自のクロールングによる収集、そしてTorのネットワークサービスディスクリプタから取得した46,562のTorアドレスと2,687のI2Pアドレスのデータを分析した。彼らは様々なデータセットを混合して分析を行ったため、データセットの収集時期についてはあいまいである。彼らの研究にはクロールングしたデータセットのカバレッジについての言及があり、彼らは研究においてカバレッジはTorネットワークの約61%と推定した。彼らの求めたカバレッジはTor Projectによって示されるHidden Serviceドメインの数を基準にした彼らの持つデータセットのドメイン数のカバレッジであった。彼らはHidden Serviceドメインの全てでhttpサーバーが動作しているとは限らない点と、ドメインがいつでもオフラインになる可能性があることを理由に、Torネットワークの正確なカバレッジを推定することは困難であると述べている。

ここで挙げた研究は、ウェブやダークウェブのノード数について言及があり、本章の取り扱うテーマと関係がある。

第3節 手法

観測されたアドレスから Tor ネットワークのサイズを推定する。人口生態学における Petersen モデルの標識再捕獲法 [55] を使用して、ネットワークノードのサイズを推定した。コンピュータネットワーク上におけるクロールのモデルは、現実世界における植物や動物の採集とよく似ている。したがって、実際の動植物の数を推定するのと同じ方法で、ネットワーク内のノードの数を推定することができる。Petersen モデルの標識再捕獲法は、次の式を使用して計算できる。

$$N = \frac{MC}{R} \quad (3)$$

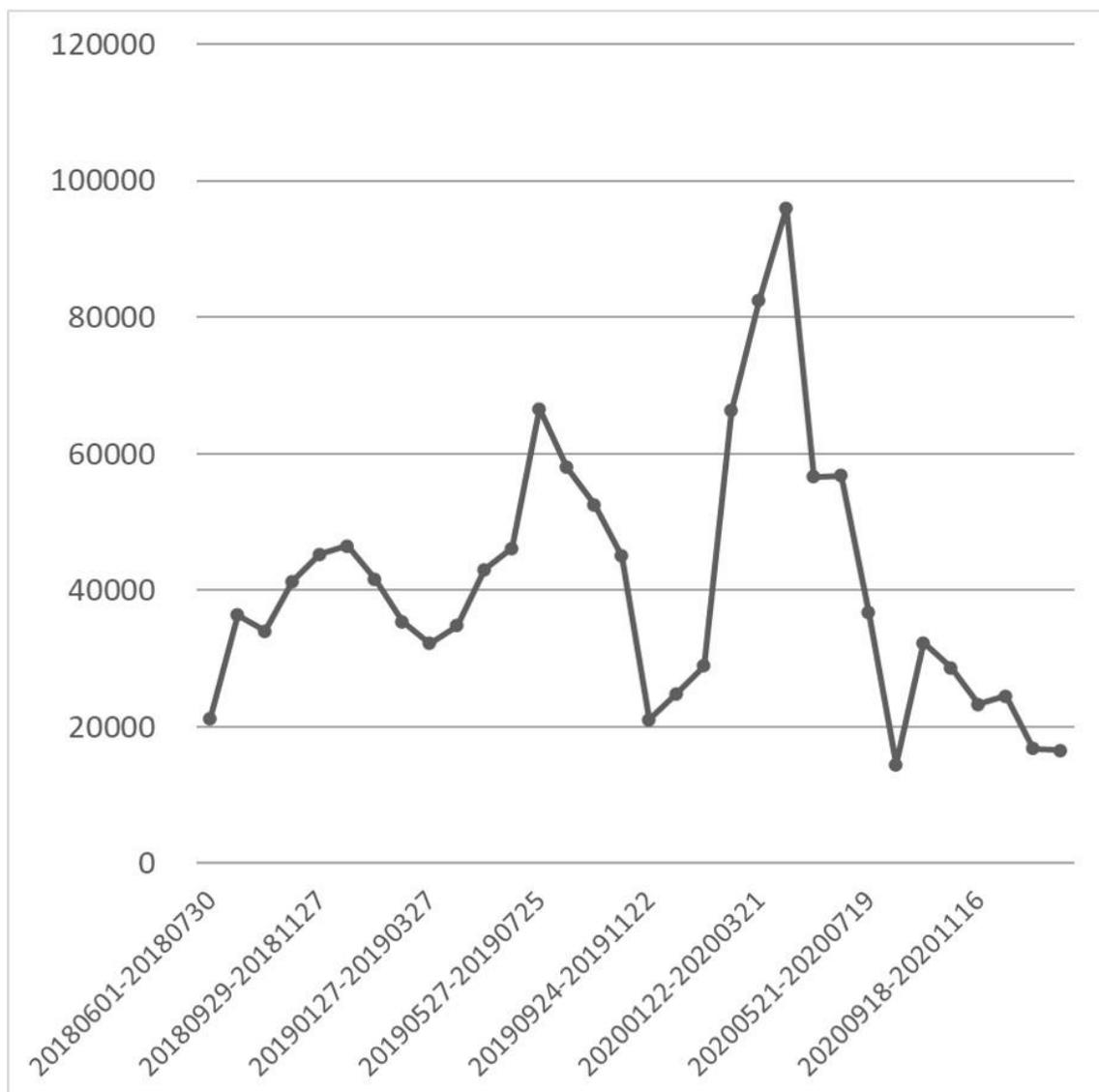
ここで、 M は最初に収集された一意のドメインの数、 C は2回目に収集された一意のドメインの数、 R は M と C を比較したときに重複する一意のドメインの数を表し、 N が、ドメインの推定総数となる。

Petersen モデルは、母集団を構成する個体数を推定するための基本的な方法であり、簡単な計算と処理によって求められる。このモデルの要件は、(1) 閉鎖系であることと、(2) 捕獲された個体が再捕獲される前に系の中で十分に混合されている必要がある。このモデルをネットワーククロールに適用する場合、捕獲対象の攪拌は、すでに収集されている対象リストから一様ランダムに選択することで容易に実現できる。閉鎖系の状態は、再捕獲サイクルを短くすることで近似する。標識はドメイン名により識別した。クロールされたデータについて、サンプリング幅 30 日ごとに観察されたドメインを集約し、最初のキャプチャと 2 番目のキャプチャの両方に存在する一意のドメインの数を計算し、推定値を求めた。

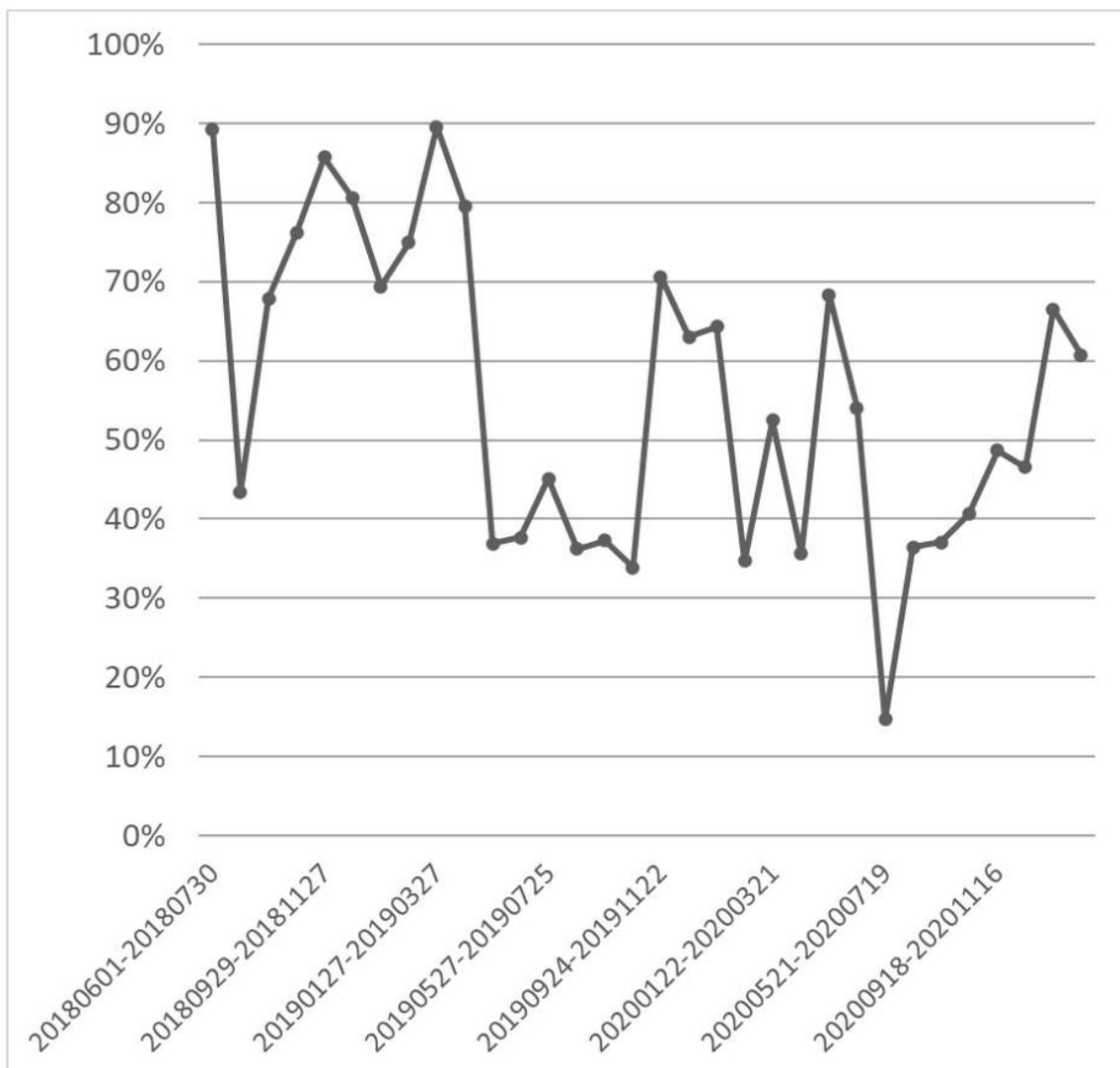
第4節 結果

サンプリング期間中に収集された一意のホストの数は、観測したデータセットの 30 日毎のサンプリングビンの集計の結果から 10^4 の付近を維持し続けている (図表 7)。ここから抽出された、各サンプリングビンに含まれる一意のドメインの数に基づいて、標識再捕獲法を使用して推定ドメイン数を計算した (図表 46)。得られた推定値は 14,509 から 96,034 までの値を示した。データセット全体の平均は 40,848 ドメインであった。標識再捕獲法の再捕獲率を図表 47 にプロットした。測定期間中の再捕獲率は 17 % から 90 % まで変

化し、平均再捕獲率は 56 %であった。再捕獲率は、収集期間を通じて徐々に低下する傾向がみられた。



図表 46. 標識再捕獲法により推定されたダークウェブのドメイン数



図表 47. 標識再捕獲法における再捕獲率

第5節 評価

標識再捕獲法により導出されたダークウェブの大きさの推定値について評価を行う。評価方法は Kitada らによる実験 [56] を参考に評価した。Peterson による標識再捕獲法 [55] は、(1) 閉鎖系であることと、(2) 捕獲された個体が再捕獲される前に十分混合されていること、を仮定においている。精度の良い推定のためには、これらの仮定が成り立っている必要がある。このモデルとネットワークのクローリングに適用する場合、(2) の再捕獲対象の攪拌については、既に収集されている収集先 URL のリストから一様ランダムに選出することで実現できる。(1) の閉鎖系については、再捕獲までの間隔を短くすることで近似する。データセット全体においてダークウェブの生存日数の中央値は 33 日であることから、その数値より小さくなると望ましい。しかし、再捕獲までの間隔が短くなると、クローリング性能に限界があるため捕獲数が減少してしまう。捕獲数が減少し、再捕獲率が小さくなると、推定誤差が大きくなってしまうため、再捕獲率はある程度高くなることが望ましい。仮に 100% の捕獲を行うことが出来れば、母集団の全てを取得したことになり、これは、ダークウェブの全てを収集したと言い換えることが出来る。本実験では 30 日幅で集計を行った結果が示され (図表 46)、測定期間中の再捕獲率は 17% から 90% まで変化し、平均再捕獲率は 56% であったことを示した (図表 47)。

Peterson モデルにおいて母集団の捕獲率 p は

$$p = \frac{M}{N} \quad (4)$$

で求まる。 M は 1 回目の捕獲数、 N は推定数をそれぞれ表す。

この推定量に対して、2 回目の捕獲数である C と再捕獲数 R を使って求まる標本標識率 \hat{p} を、

$$\hat{p} = \frac{R}{C} \quad (5)$$

とすると、標本個体数 \hat{N} は

$$\hat{N} = \frac{M}{\hat{p}} = \frac{MC}{R} \quad (6)$$

で推定され、 \hat{N} の近似的な分散は

$$\hat{V}(\hat{N}) = M^2 \hat{V}\left(\frac{1}{\hat{p}}\right) \cong \frac{\hat{N} - C}{\hat{N} - 1} \cdot \frac{M^2(1 - \hat{p})}{C\hat{p}^3} = \frac{M^2 C^2 (M - R)(C - R)}{(MC - R)R^3} \quad (7)$$

で推定される。個体数が多い場合は有限修正項が無視され

$$\hat{V}(\hat{N}) \cong \frac{M^2 C^2 (C - R)}{R^3} \quad (8)$$

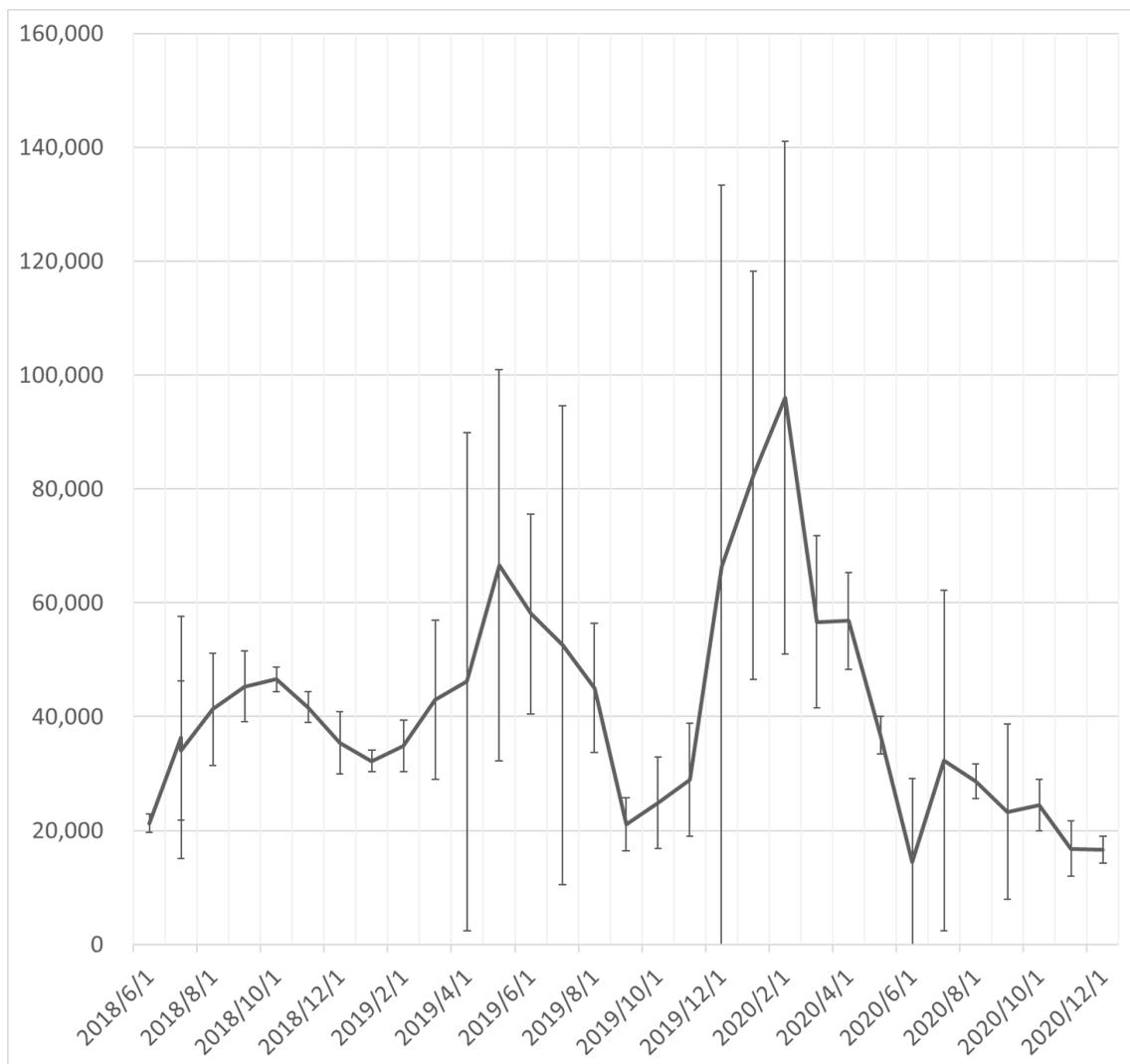
と簡略化できる。変動係数は

$$V = \sqrt{\frac{(1 - \hat{p})}{n\hat{p}}} \quad (9)$$

で示される。母集団から得られた標本の標本平均からの誤差は中心極限定理から母集団の分布に抛らず正規分布と仮定できるため、本実験データに対して得られた分散から標準偏差を求め、推定値とともにプロットすると図表 48 及び図表 49 が得られる。

結果からは、精度よく推定できている時期と、精度が低く推定されている時期が現れた。推定値の変動の少ない時期である 2018 年 10 月頃や 2019 年 10 月頃、2020 年 10 月頃は標準偏差が 2,000 前後と少なく、推定値が大きく変動する 2019 年 4 月や 2019 年 12 月には標準偏差が 40,000~100,000 近くになることもあり、大きいことが示された。再捕獲率をプロットした図表 47 と合わせて見てみると、再捕獲率が高いことだけが推定精度を高める要因とは必ずしもいえず、誤差が大きくなる 2019 年 4 月や 2019 年 11 月のタイミングで再捕獲率の大きな変動が起こっていることが読み取れる。本実験における推定値 N の変動係数はいずれも Kitada ら [56] により示されるものと比べて小さな値を示した。Peterson モデルの推定手法は閉鎖系であることを仮定においているため、ダークウェブに急激な変化があることでこの仮定が崩れ、推定精度を低くする要因になっているものと考えられる。

評価の結果、本実験によって得られたダークウェブの大きさの推定値の内、2018 年 9 月頃~2019 年 2 月頃、2019 年 10 月頃、2020 年 4 月頃、2020 年 10 月~1 月頃の推定値については精度良く推定できているといえる。それ以外の時期については、大きな誤差が含まれている点を加味して読み解く必要がある。



図表 48. 本実験におけるダークウェブの大きさの推定値と誤差の範囲

図表 49. 本実験におけるダークウェブの大きさの推定に使用した値

Date	1回目の 捕獲数 (M)	2回目の 捕獲数 (C)	再捕獲数 (R)	推定値 (N)	標準偏差	変動係数
2018/6/1	19,701	20,458	18,943	21,277	1,637	0.002
2018/7/1	20,458	23,871	13,420	36,390	21,248	0.008
2018/7/31	23,871	32,549	22,817	34,053	12,161	0.004
2018/8/30	32,549	39,569	31,171	41,318	9,880	0.003
2018/9/29	39,569	44,388	38,752	45,324	6,159	0.002
2018/10/29	44,388	39,076	37,265	46,545	2,209	0.003
2018/11/28	39,076	30,344	28,449	41,679	2,688	0.004
2018/12/28	30,344	30,346	25,997	35,420	5,484	0.003
2019/1/27	30,346	30,580	28,793	32,229	1,941	0.002
2019/2/26	30,580	31,141	27,347	34,823	4,527	0.003
2019/3/28	31,141	18,484	13,381	43,017	13,958	0.011
2019/4/27	18,484	39,980	16,010	46,158	43,732	0.007
2019/5/27	39,980	42,956	25,792	66,586	34,336	0.006
2019/6/26	42,956	24,094	17,832	58,041	17,534	0.010
2019/7/26	24,094	35,072	16,077	52,561	42,046	0.008
2019/8/25	35,072	16,890	13,151	45,043	11,300	0.012
2019/9/24	16,890	18,010	14,449	21,053	4,647	0.005
2019/10/24	18,010	20,592	14,925	24,848	8,032	0.006
2019/11/23	20,592	25,149	17,906	28,921	9,871	0.005
2019/12/23	25,149	53,512	20,272	66,386	66,998	0.007
2020/1/22	53,512	60,365	39,195	82,415	35,869	0.004
2020/2/21	60,365	43,382	27,269	96,034	44,990	0.008
2020/3/22	43,382	48,940	37,483	56,642	15,152	0.003
2020/4/21	48,940	33,624	28,970	56,802	8,470	0.005
2020/5/21	33,624	5,483	5,011	36,791	3,313	0.034
2020/6/20	5,483	13,220	4,996	14,509	14,682	0.012
2020/7/20	13,220	25,791	10,548	32,324	29,873	0.009
2020/8/19	25,791	12,205	10,973	28,687	3,054	0.012
2020/9/18	12,205	20,384	10,674	23,308	15,343	0.008
2020/10/18	20,384	12,606	10,484	24,510	4,524	0.010
2020/11/17	12,606	14,399	10,786	16,829	4,879	0.006
2020/12/17	14,399	11,121	9,643	16,606	2,370	0.008

第6節 議論

ダークウェブのサイズを測定することは難しいと考えられている [28]. 本研究では, 個体群生態学で一般的な標識再捕獲法を使用して, 観測データからダークウェブノードの数を推定した (図表 46). 本研究の推定では, さまざまな時点で 32 回の測定を行い, 14,509 から 96,034 の推定値が示された. データセット全体の推定値平均は 40,848 であった. 標識再捕獲法の実用性に関する以前の研究 [56]によると, 標識再捕獲率が 10 % の場合, 相対誤差は ± 0.5 の範囲にあることが示されている. 本研究の観察データでは, 平均再捕獲率は 56 % であり, この期間中も 10% より高い再捕獲率が継続的に維持されていた (図表 47). クロールターゲットは, 収集されたすべての候補となるターゲットから一様ランダムに選択されているため, 本研究で観察したデータには大きな偏りがないことが視覚化した結果にも示されている (図表 12). 閉鎖系に近づけるために, 再捕獲期間はスナップショット幅よりも短い 30 日間隔に設定された. 標識として使用されるドメイン名は, ハッシュの持つ衝突困難性から, 攻撃者によって意図的に再取得されることは計算量的に困難である. したがって, ドメイン名は標識として正しく機能することが期待できる.

第7節 本章のまとめ

本研究では, 個体群生態学で一般的な標識再捕獲法を使用して, 観測データからダークウェブノードの数を推定した. 本研究の推定では, 観測期間中のさまざまな時点で 32 回の測定を行い, ダークウェブの大きさは 14,509 から 96,034 であるとの推定値が示された. データセット全体から求まる推定値平均は 40,848 であった. 評価の結果, 求まったダークウェブの大きさの推定値の内, 2018 年 9 月頃~2019 年 2 月頃, 2019 年 10 月頃, 2020 年 4 月頃, 2020 年 10 月~1 月頃の推定値については他の時期と比べ, 精度良く推定できていることが分かった.

第6章

結論

本研究では、ダークウェブのクロールデータから半年の期間に分割されたスナップショットを作成した。時系列グラフを使用して、ダークウェブネットワークの経時変化を調査した。

第3章では、ダークウェブの探索について記載した。本研究では、観測時点においてダークウェブとして人気のあるTorネットワークを対象に観測を行った。ダークウェブは、その言葉の登場にあたり、検索エンジンによる検索では発見できないウェブの領域として位置づけられているため、自身により観測手段を確保する必要がある。本研究における観測により、おおよそ2年に渡る2018年6月1日から2021年1月30日までの期間において、25,270,157ページのhtmlテキストファイルを収集した。これらのテキストファイルは、幅優先探索によりTorネットワークから収集された。これらのテキストファイルからハイパーリンクを抽出し、グラフデータを構築した。データセットは180収集日の時系列で分割し、分析のためのダークウェブのスナップショットとした。これらのスナップショット間において、ドメインテキストの差分の視覚化を行い、ダークウェブのドメイン名の分布や生成・消失について偏りがみられないかを調べた。更に、ダークウェブのドメインの生存期間を調べ、統計情報を提供する。本章の結果として、ダークウェブのネットワークはドメイン名の観点から偏りは確認できず、適正な状態であることが確認できた。そして、ダークウェブの生存期間から求まるダークウェブの半減期についてもワールドワイドウェブと比べて1.5%~2.3%しかなく、著しく短いことが示された。

第4章では、ダークウェブネットワークの視覚化について記載した。第3章で作成したスナップショットの有向グラフデータについてばねモデルを用いて視覚化を行った。視覚化において、エッジの始点と終点でそれぞれ異なる色をつけることで、本研究特有の視覚化効果を得た。この視覚効果からは、ダークウェブグラフにおけるハブノードの存在と、低い次数で多くのノードが結合する領域の存在を浮き彫りにした。これらの視覚化結果を、ネットワークグラフの代表的な中心性指標と比較を行った結果、中心性指標からは特徴があるとみなされないいくつかの領域を視覚化により発見することができた。更に、これら

の視覚化結果を時系列で比較した結果、ダークウェブにおける情報検索の在り方が、リンクコレクションを中心とした関係性をもつ構造から、検索エンジンと相互参照を中心とする関係性をもつ構造へと徐々に変化している様子を示した。視覚化による領域からは、ダークウェブの中でも大きな勢力を持つポルノネットワークやフィッシングの存在が確認できた。分析の結果、これらの勢力は観測を開始した当初は存在せず、2019年5月以降に誕生したことが明らかとなった。

第5章では、ダークウェブネットワークのノード数の推定について記載した。ダークウェブの大きさを求めることは困難であると考えられており、本研究においてもダークウェブの全てを瞬時に観測することは実現できていない。そこで、本研究では観測データをもとに、Petersonによる標識再捕獲法を用いてダークウェブの大きさの推定を試みた。Petersonによる標識再捕獲法は個体群生態学の分野で野生動植物の個体数の推定を行う場合にしばしば使用される手法である。野生動植物の個体数も、あるフィールドにおいて全ての個体を同時に識別してカウントすることは困難であるため、ダークウェブの大きさの推定と同質の問題を抱えている。本研究の推定では、観測期間中のさまざまな時点で32回の測定を行い、ダークウェブの大きさは14,509から96,034であるとの推定値が示された。データセット全体の推定値平均は40,848であった。

本研究が視覚的にダークウェブの変化を示すまで、ダークウェブがどのように変化してきたかを知る手段は無かった。本研究が実現した視覚化は、ダークウェブという巨大なネットワークがどのように動き、変化してきたかを描写し、過去と現在のダークウェブの姿、そして将来のダークウェブの変化についての洞察を得るための手段を提供する。本研究の観測期間において起こったダークウェブの最大の変化は、ダークウェブの中の情報検索のあり方が、リンク集から検索エンジンへ変化しつつあることを示唆する結果を得たことである。また、視覚的な情報からは、ネットワーク中心性指標では見えづらい領域の存在を発見できることもわかった。これらの領域は少ないリンクによってお互いが弱く結合するため、次数をもとにランクを求めるネットワーク中心性指標では数値が低くなっていた。

本研究の貢献は、直感的で理解しやすい視覚化を通じて、ダークウェブが時間の経過とともにどのように変化したかを示したことにある。これは、ダークウェブの変化を理解するために重要な情報である。

本研究のデータは、クローリングによってTorネットワークのハイパーリンクから追跡できる領域のみを反映している。今後の作業では、httpプロトコル以外のネットワークや、ハイパーリンクからたどることができないノードを含めると、新しい発見につながる可能性がある。本研究の分析では、180収集日のデータに分割されたスナップショットを使用している。この期間を増減することで、将来的にはさまざまな洞察を得ることができる可能性がある。また、本研究の視覚化は、計算の複雑さを軽減するためにドメインの関係のみを示している。これをページ単位にすることで、より解像度の高い変化について洞察を得られる可能性がある。ドメインの出現や消失に着目した分析を行うことや、その変化が見

えるよう映像化や3次元化、座標情報を利用したオーバーレイによる視覚的なマッピング、そしてインタラクティブな操作による変化などの視覚的な表現を模索することも変化に対する洞察を得る上で興味深い。本研究において視覚的に明らかとなった不正なネットワークについて自動的に抽出する手法の検討も効率性の点から有益である。

我々の知る限り、ダークウェブのネットワークは巨大で複雑でありそして流動的であるため、これまで簡単に理解することはできなかった。本研究の貢献は、ダークウェブネットワークの成長に対する直感的な理解と洞察を提供したことにある。ダークウェブネットワークの変化から明らかとなったこれらの洞察は、ダークウェブに対する理解を深めるとともに、サイバーセキュリティに関連するさまざまな防衛戦略の検討に役立つものと期待している。

参考文献

- [1] Taichi Aoki, et al., “Survey on attack methods of anonymity communication systems,” Proceedings of Computer Security Symposium 2015, 2015.3 (2015): 1207-1212.
- [2] Andy (26 November 2009) Beckett, *The dark side of the internet*, <https://web.archive.org/web/20130908073158/http://www.theguardian.com/technology/2009/nov/26/dark-side-internet-freenet>, December 2 2021.
- [3] Tor, *The tor network*, <http://www.torproject.org/>, December 2 2021.
- [4] I2P, *The i2p network*, <http://www.i2p2.de/>, December 2 2021.
- [5] Wikipedia, *Dream Market*, https://en.wikipedia.org/wiki/Dream_Market, December 2 2021.
- [6] Darknetlive, *Gangsta ’ s Paradise Forum Shut Down by Dread*, <https://darknetlive.com/post/gangsta-s-paradise-forum-shut-down-by-dread/>.
- [7] Department of Justice, *Three Germans Who Allegedly Operated Dark Web Marketplace with Over 1 Million Users Face U.S. Narcotics and Money Laundering Charges*, <https://www.justice.gov/opa/pr/three-germans-who-allegedly-operated-dark-web-marketplace-over-1hyphenmillion-users-face-us>, December 2 2021.
- [8] EUROPOL, *DOUBLE BLOW TO DARK WEB MARKETPLACES*, <https://www.europol.europa.eu/newsroom/news/double-blow-to-dark-web-marketplaces>, December 2 2021.
- [9] EUROPOL, *DEEPDOTWEB SHUT DOWN: ADMINISTRATORS SUSPECTED OF RECEIVING MILLIONS OF KICKBACKS FROM ILLEGAL DARK WEB PROCEEDS*, <https://www.europol.europa.eu/newsroom/news/deepdotweb-shut-down-administrators-suspected-of-hyphenreceiving-millions-of-kickbacks-illegal-dark-web-proceeds>, December 2 2021.
- [10] TechNadu, *“Dark Web News” Discontinued in Fear of FBI and Europol Arrests*, <https://www.technadu.com/dark-web-news-discontinued-fear-fbi-europol-arrests/67252/>, December 2 2021.
- [11] EUROPOL, *MULTI-MILLION EURO CRYPTOCURRENCY LAUNDERING*

- SERVICE* *BESTMIXER.IO* *TAKEN* *DOWN*,
<https://www.europol.europa.eu/newsroom/news/multi-million-euro-cryptocurrency-laundering-service-bestmixerio-taken-down>, December 2 2021.
- [12] MhdWesam Al-Nabki, Eduardo Fidalgo, Enrique Alegre, and Laura Fernández-Robles, “Torank: Identifying the most influential suspicious domains in the tor network,” *Expert Systems with Applications*, 123:212–226, 2019.
- [13] Arnaud Sallaberry, Faraz Zaidi, Christian Pich, and Guy Melançon, “Interactive visualization and navigation of web search results revealing community structures and bridges,” In *Proceedings of Graphics Interface*, pages 105–112, 2010.
- [14] Masashi Toyoda and Masaru Kitsuregawa, “A system for visualizing and analyzing the evolution of the web with a time series of graphs,” In *Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 151–160, 2005.
- [15] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener, “Graph structure in the web,” *Computer networks*, 33(1-6):309–320, 2000.
- [16] Jon M Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S Tomkins, “The web as a graph: Measurements, models, and methods,” In *International Computing and Combinatorics Conference*, pages 1–17. Springer, 1999.
- [17] Ruslan Enikeev, *The Internet map*, <https://internet-map.net/>, December 2 2021.
- [18] Yilu Zhou, Edna Reid, Jialun Qin, Hsinchun Chen, and Guanpi Lai, “Us domestic extremist groups on the web: link and content analysis,” *IEEE intelligent systems*, 20(5):44–51, 2005.
- [19] Tianjun Fu, Ahmed Abbasi, and Hsinchun Chen, “A focused crawler for dark web forums,” *Journal of the American Society for Information Science and Technology*, 61(6):1213–1231, 2010.
- [20] Christopher C Yang and Tobun D Ng, “Terrorism and crime related weblog social network: Link, content analysis and information visualization,” In *2007 IEEE Intelligence and Security Informatics*, pages 55–58. IEEE, 2007.
- [21] Christopher C Yang, Nan Liu, and Marc Sageman, “Analyzing the terrorist social networks with visualization tools,” In *IEEE Symposium on Information Visualization*, 2005. INFOVIS 2005., pages 32–39. IEEE, 2005.
- [22] Georgia Avarikioti, Roman Brunner, Aggelos Kiayias, Roger Wattenhofer, and Dionysis Zindros, “Structure and content of the visible darknet,” *arXiv preprint arXiv:1811.01348*, 2018.

- [23] Virgil Griffith, Yang Xu, and Carlo Ratti, “Graph theoretic properties of the darkweb,” arXiv preprint arXiv:1704.07525, 2017.
- [24] Manlio De Domenico and Alex Arenas, “Modeling structure and resilience of the dark network,” *Physical Review E*, 95(2):022313, 2017.
- [25] WooHyun Park, “A study on analytical visualization of deep web,” In 2020 22nd International Conference on Advanced Communication Technology (ICACT), pages 81–83. IEEE, 2020.
- [26] Gast´on L’huillier, Hector Alvarez, Sebasti´an A R´ıos, and Felipe Aguilera, “Topic-based social network analysis for virtual communities of interests in the dark web,” *ACM SIGKDD Explorations Newsletter*, 12(2):66–73, 2011.
- [27] Jeffrey Heer and Danah Boyd, “Vizster: Visualizing online social networks,” In IEEE Symposium on Information Visualization, 2005. INFOVIS 2005., pages 32–39. IEEE, 2005.
- [28] Carlos Cilleruelo, Luis De-Marcos, Javier Junquera-Sanchez, and Jose-Javier Martinez-Herraiz, “Interconnection between darknets,” *IEEE Internet Computing*, 2020.
- [29] Hyperion Gray, *Dark Web Map v2*, <https://www.hyperiongray.com/dark-web-map/>, December 2 2021.
- [30] Taichi Aoki and Atsuhiko Goto, “Graph visualization of the dark web hyperlink,” In 2020 The Eighth International Symposium on Computing and Networking (CANDAR). IEEE, 2020.
- [31] Taichi Aoki and Atsuhiko Goto, “Graph visualization of dark web hyperlinks and their feature analysis,” *International Journal of Networking and Computing* 11.2 (2021): 354-382..
- [32] Juha Nurmi, *Ahmia*, <https://ahmia.fi/>, December 2 2021.
- [33] Loesing, Karsten, Steven J. Murdoch, and Roger Dingledine, “A case study on measuring statistical data in the Tor anonymity network,” *International Conference on Financial Cryptography and Data Security*. Springer, Berlin, Heidelberg, 2010.
- [34] DIFF, *WinMerge*, <https://winmerge.org/>, December 2 2021.
- [35] Unperson Hiro, *Eschalot*, <https://github.com/ReclaimYourPrivacy/eschalot>, December 2 2021.
- [36] Yosuke Miyata, et al., “Web Page Life: A Survey of 10 Million Pages Existing in 2001,” *Collection of papers presented at the Japan Society of Library and Information Science* 61 (2013), 33-36.

- [37] Spinellis, Diomidis, “The decay and failures of web references,” *Communications of the ACM* 46.1 (2003): 71-77..
- [38] Jun Iio, “Web Decay in a Series of Technical Columns,” *IEICE Conferences Archives*. The Institute of Electronics, Information and Communication Engineers, 2011.
- [39] FBI, *JCODE Actions in Los Angeles Shut Down Major Darknet Drug Vendor*, <https://www.fbi.gov/news/stories/operation-disruptor-jcode-shuts-down-darknet-drug-vendor-092220>, December 2 2021.
- [40] EUROPOL, *DARKMARKET: WORLD ’ S LARGEST ILLEGAL DARK WEB MARKETPLACE TAKEN DOWN*, <https://www.europol.europa.eu/newsroom/news/darkmarket-worlds-largest-illegal-dark-web-marketplace-taken-down>, December 2 2021.
- [41] Fruchterman, Thomas MJ, and Edward M. Reingold, “Graph drawing by force - directed placement,” *Software: Practice and experience* 21.11 (1991): 1129-1164..
- [42] D3, *Force-directed graph layout*, <https://github.com/d3/d3-force/>, December 2 2021.
- [43] Baur, Michael, and Ulrik Brandes, “Crossing reduction in circular layouts,” *International Workshop on Graph-Theoretic Concepts in Computer Science*. Springer, Berlin, Heidelberg, 2004..
- [44] Slade, Michael, “A layout algorithm for hierarchical graphs with constraints,” (1994).
- [45] Herman, Ivan, Guy Melançon, and M. Scott Marshall, “Graph visualization and navigation in information visualization: A survey,” *IEEE Transactions on visualization and computer graphics* 6.1 (2000): 24-43..
- [46] Naoki TAKAHASHI and Takashi SAKAMOTO and Toshikazu KATO, “Color-analysis of Image Photography of Interior Brand based on Extraction of Representative Colors and Clustered Color-features,” *Transactions of Japan Society of Kansei Engineering* 15.1 (2016): 203-212..
- [47] Vincent, Luc, and Pierre Soille, “Watersheds in digital spaces: an efficient algorithm based on immersion simulations,” *IEEE Transactions on Pattern Analysis & Machine Intelligence* 13.06 (1991): 583-598.
- [48] Mordvintsev, Alexander, and K. Abi, *OpenCV-python tutorials documentation*, Obtenido de <https://media.readthedocs.org/pdf/opencv-python-tutroals/latest/opencv-python-tutroals.pdf> (2014).
- [49] Kornilov, Anton S., and Ilia V. Safonov, “An overview of watershed algorithm

- implementations in open source libraries,” *Journal of Imaging* 4.10 (2018): 123.
- [50] CMM, *IMAGE SEGMENTATION AND MATHEMATICAL MORPHOLOGY*, <https://people.cmm.minesparis.psl.eu/users/beucher/wtshed.html#watshed>.
- [51] Bang-Jensen, Jørgen, and Gregory Gutin, “Theory, algorithms and applications,” Springer Monographs in Mathematics, Springer-Verlag London Ltd., London 101 (2007).
- [52] NetworkX, *Network Analysis in Python*, <https://networkx.org/>, December 2 2021.
- [53] Sergey Brin and Lawrence Page, “The anatomy of a large-scale hypertextual web search engine,” *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.
- [54] Coffman, Kerry, and Andrew Odlyzko, “The size and growth rate of the Internet,” (1998)..
- [55] PB Best and RWRand, “Results of a pup-tagging experiment on the arctocephalus pusillus rookery at seal island, false bay, south africa,” *Rapports et Proces-Verbaux des Reunions (Denmark)*, 1975.
- [56] Shuichi Kitada, Sachio Sekiya, and Masashi Yokota, “An evaluation of petersen method through experiments in a rearing tank,” *Journals of Japanese Society of Fisheries Science*, 67(2):203–208, 2001.

謝辞

本研究の遂行にあたり，研究の進め方や方向性について日頃より熱心にご指導を頂いている後藤厚宏先生に深く感謝致します。副査を引き受けてくださった豊田正史教授をはじめ，査読を行っていただきました大久保隆夫教授，松井俊浩教授へ心より感謝致します。様々な助言や研究への励みをいただきました情報セキュリティ大学院大学の皆様に重ねて感謝致します。本研究の活用と多くの助言を頂きました一般財団法人日本サイバー犯罪対策センターの坂明氏をはじめ，関わった多くの皆様に深く感謝いたします。

研究業績

査読付学術論文

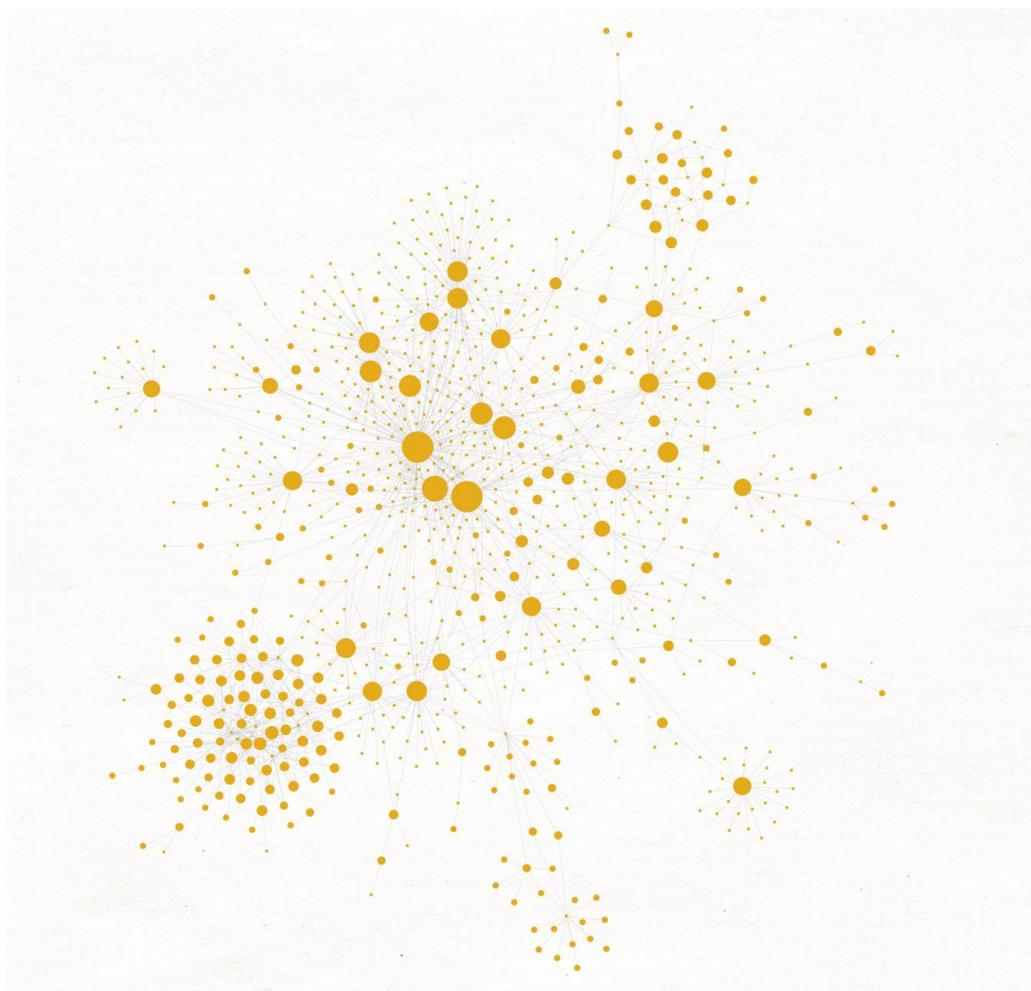
AOKI, Taichi; GOTO, Atsuhiko. Graph visualization of dark web hyperlinks and their feature analysis. *International Journal of Networking and Computing*, 2021, 11.2: 354-382.

査読付国際会議（プロシーディングとして採録）

AOKI, Taichi; GOTO, Atsuhiko. Graph visualization of the dark web hyperlink. In: 2020 Eighth International Symposium on Computing and Networking (CANDAR). IEEE, 2020. p. 89-94.

付録 A

ここでは、代表的なグラフィレイアウトの描画イメージとして、ばねモデルによるもの（図表 50）、サークルモデルによるもの（図表 51）、階層モデルによるもの（図表 52）（図表 53）、の3つを掲載する。いずれの描画モデルもグラフ構造を自動的にレイアウトする手法であり、それぞれメリットとデメリットがある。



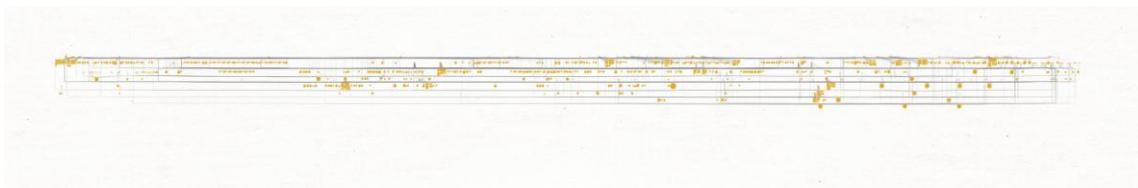
図表 50. ばねモデルによるグラフィレイアウトのイメージ

力学モデルは、描画がコンパクトにまとまるため、全体像が把握しやすい。



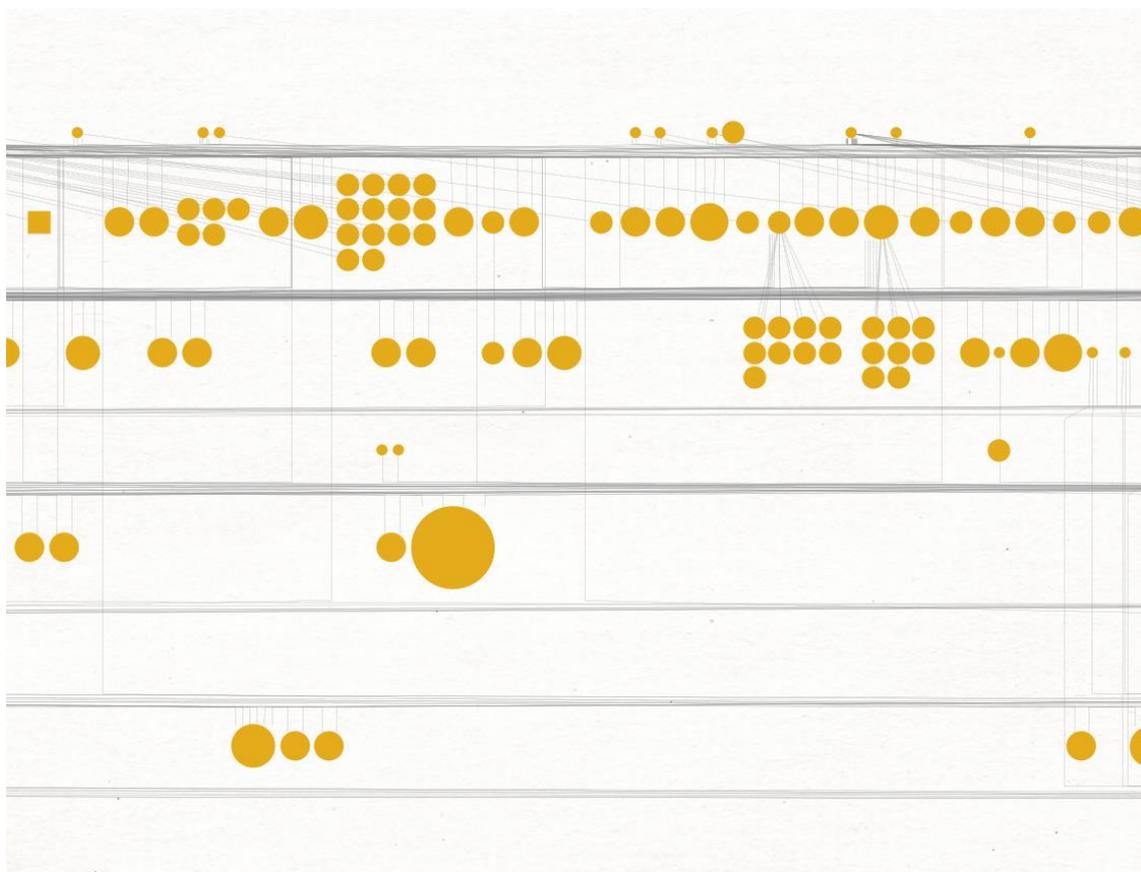
図表 51. サークルモデルによるグラフレイアウトのイメージ.

サークルモデルは、ノード間の結合関係を見る場合に適している.



図表 52. 階層モデルによるグラフィレイアウトのイメージ

階層モデルは、階層関係を見る場合に適している。



図表 53. 階層モデルによるグラフィレイアウトの拡大図

付録 B

本稿で使用した各スナップショットにおけるネットワーク中心性指標を掲載する。

付録 B 1. 入力次数による中心性指標一覧

図表 54. スナップショット A の入力次数の上位 10 ノード (n=48360)

rank	node	in-degree	centra	title
1	dhosting4okcs22v.onion	0.015		Daniel's Hosting
2	blockchainbdgpzk.onion	0.005		Blockchain
3	lchudifyeqm4ldjj.onion	0.004		Dream Market
4	tmskhzavkydupbr.onion	0.003		Dream Market
5	jd6yhuwcivehvd4.onion	0.003		Dream Market
6	blockchainbdgpz.onion	0.003		Blockchain
7	6qlocfg6zq2kyacl.onion	0.003		Dream Market
8	k3pd243s57ftnpa.onion	0.003		Dream Market
9	7ep7acrkunzdcw3l.onion	0.003		Dream Market
10	xytjqcfendzeby22.onion	0.003		Proof of onion url correctness

図表 55. スナップショット B の入力次数の上位 10 ノード (n=53300)

rank	node	in-degree	centra	title
1	dhosting4xxoydyavckq7tsmtgi4wfs3flpeyitekkmqwu4v4r46syd.	0.005853768		Daniel's Hosting - Info
2	blockchainbdgpzk.onion	0.003377174		Blockchain
3	dc72oitx4plhel3r.onion	0.003283364		Hidden Wiki The Directory Of New Updated Onion Links
4	torpress2sarn7xw.onion	0.002720501		REZNOV
5	z5taguvepvuevyp3.onion	0.002382784		Phishing
6	au54gdcynh7jvet6.onion	0.002251449		VendorPro - Best PayPal & Bank Account Vendor On Tor
7	xpotbpgfnliidudm.onion	0.002157639		Black Label Kush 10G - Pot Shop
8	torbox3uiot6wchz.onion	0.001819922		TorBox
9	blockchainbdgpz.onion	0.001801159		Blockchain
10	lchudifyeqm4ldjj.onion	0.001688587		Dream Market

図表 56. スナップショット C の入力次数の上位 10 ノード (n=70684)

rank	node	in-degree	centra	title
1	cashvd5pznwrucio.onion	0.003		Cash Machine For Everybody - Easy to use - Paypal Account - Paysafecard - US and EU Credit Cards
2	r72jijnwl6nh47zaks2loja4i34gh4soytlpk2tzsnd2s4rjptxf6ayd.on	0.003		Zoophilia Horse _ Golden Club Hugry For Kids _ Onion Wiki
3	3h4ctuedj6meudw2k2u3td4j472rwqxeqvknkou7bvfrcwztt66gfi	0.003		Zoo Sex Free Golden Club Community - Tor Forbidden Fruit TOR List
4	dhosting4xxoydaivckq7tsmtgi4wfs3flpeyitekkmqwu4v4r46syd.	0.003		Daniel's Hosting - Info
5	oqqsn7yplyu24dfirwerruz6bdupi6gqxzdxoaoacteyzoyfnyhxn7qd.c	0.003		free porn youtube video
6	nj3xpw7vvzca3mvvaagzeri6nqclvdrdc5jdcc6t34q2hsiiq7dchad.	0.003		We Video Collection # Boys Town # Verified Onion
7	xq4qjtl354vrst67ko3dait6mmw3r4nbzry332lqrivyz4jwvsv3ogad.	0.003		Zoophilia Horse _ Golden Club Hugry For Kids _ Onion Wiki
8	uyqmqzquv252txhkqfy54a5bn7qk2z3v4vargjvqzm6wwrvgn6lod3f	0.003		Young Legal Porn E * Grand Mallu Porn Video As Well As Kazashki Porno Video Adjusted Elevator Porn Video
9	hueonvufzh2pps3gfsistijp5u2v5hrohci4gc27venko2thuv63yad.c	0.003		Sites Child Porn Preview Fre # 4chan Child Porno # Secure Onions
10	i3eauqwhnya2tdkxmp4enn4wi6h4mhu5nivkruwtet7qhd7mffi4c	0.003		Young Teen Ass _ 1st Siberian Mouse _ Porn Onions

図表 57. スナップショット D の入力次数の上位 10 ノード (n=103674)

rank	node	in-degree	centra	title
1	oa7afrpjhd6ffocc.onion	0.014		n/a
2	m6kshkn2vcovss35.onion	0.012		Reaper Mafia
3	z5taguvepvuevyp3.onion	0.010		Phishing
4	2pneiouz2aj27kjs.onion	0.009		VendorPal - Buy US Paypal \$50.69
5	qyh35wx75lkf6z55.onion	0.009		The Way it works
6	sk3w2x7g2gksov6.onion	0.009		MILLIONAIR Private Club
7	ahhq2dssfo7e77ui.onion	0.006		n/a
8	lolipornp7zsenhy3k3khowbdhyn4afshaj6axw7lakg5xgguezcinqd	0.006		LoliPorn
9	lolipornngyulcxe3goplk3ciao65lpxfhgkph6fdompotzjxlkplmid.oni	0.006		LoliPorn
10	loliporn5lm2t3lvom6k7jjhduhpsty43g7izoerx475jtuticpiqd.onic	0.006		LoliPorn

図表 58. スナップショット E の入力次数の上位 10 ノード (n=54509)

rank	node	in-degree	centra	title
1	oa7afrpjhd6ffocc.onion	0.035		n/a
2	m6kshkn2vcovss35.onion	0.030		Reaper Mafia
3	qyh35wx75lkf6z55.onion	0.024		The Way it works
4	2pneiouz2aj27kjs.onion	0.024		VendorPal - Buy US Paypal \$50.69
5	sk3w2x7g2gksov6.onion	0.024		MILLIONAIR Private Club
6	z5taguvepvuevyp3.onion	0.022		Phishing
7	coud657j6pcf6z.onion	0.017		PenguinX - Coin Mixing Website
8	lolipornngyulcxe3goplk3ciao65lpxfhgkph6fdompotzjxlkplmid.oni	0.017		LoliPorn
9	lolipornp7zsenhy3k3khowbdhyn4afshaj6axw7lakg5xgguezcinqd	0.016		LoliPorn
10	lolipornctvs4f7k6rnr4rqutnsbjvbrpwyb6nzhcva6tmvuw34yd	0.016		LoliPorn

図表 59. スナップショット F の入力次数の上位 10 ノード (n=17464)

rank	node	in-degree	centra	title
1	loliporns6s6qmtwejbh3x7vo35qarn3msf2tjnyxk3pyd7nqrz5ecac	0.045		LoliPorn
2	lolipornngyulcxe3goplk3cjao65lpxfhgkph6fdompotzjxlkplmid.oni	0.045		LoliPorn
3	loliporns5lm2t3lvojm6k7jjhduhpsty43g7izoerx475jtuticipqd.onic	0.045		LoliPorn
4	loliporndhbnjnw3agvqhmgz5rdeic2thqbz12zu7v4ny2oiziop6ad.o	0.045		LoliPorn
5	lolipornp7zsenhy3k3khowbdhyn4afshaj6axw7lakg5xgguezcinqd	0.045		LoliPorn
6	lolipornkmcw26upm5zlh24nsnm6je6xeixzn5yacx7te3mx34xxqi	0.045		LoliPorn
7	lolipornmjc3kwke6hpawsdn44it72k7si273wtkegxvqti2m5ef2har	0.045		LoliPorn
8	lolipornctvs4f7k6rnr4rqutnsbjvbrpwyb6nzhcvxa6tmvuwh34yd	0.044		LoliPorn
9	lolipornnsman35yfd4kxabjpw6t37xs67iogwe3ufxkgykii6queyd.	0.044		LoliPorn
10	video444iwyf5364.onion	0.043		Child Porno Studios Video Store High Quality Child Porno Unlimited Download icloud leak

図表 60. スナップショット A-F の入力次数の上位 10 ノード (n=172740)

rank	node	in-degree	centra	title
1	oa7afrpjhd6ffocc.onion	0.014		n/a
2	m6kshkn2vcovss35.onion	0.012		Reaper Mafia
3	qyh35wx75lkf6z55.onion	0.010		The Way it works
4	2pneiouz2aj27kjs.onion	0.010		VendorPal - Buy US Paypal \$50.69
5	sk3w2x7g2gksovk6.onion	0.010		MILLIONAIR Private Club
6	z5taguvepvuevyp3.onion	0.008		Phishing
7	ahhq2dssfo7e77ui.onion	0.007		n/a
8	blockchainbdgpkz.onion	0.006		Blockchain
9	cpmovieswo2ww7346l6qydxeg7vbeoey3hk23uh4ivdvdb5w7lpw.	0.006		CP Video Store Pedo Videos for Bitcoin! Unlimited Download german pedo chat
10	lolipornngyulcxe3goplk3cjao65lpxfhgkph6fdompotzjxlkplmid.oni	0.006		LoliPorn

付録 B 2. 出力次数による中心性指標一覧

図表 61. スナップショット A の出力次数の上位 10 ノード (n=48360)

rank	node	degree	centrality	title
1	56wr4dvq3abd2ivkf5z36nortvu7dgona55zqsihfaqo2aeg5er4moi	0.602		Fresh Onions
2	zlal32teyptf4tvi.onion	0.532		Fresh Onions
3	msydqstlz2kzerdg.onion	0.462		Ahmia
4	msydqstjd6vtexpq.onion	0.444		Ahmia
5	msydqstlzbk3tr5q.onion	0.434		Ahmia
6	wrrkz262g55scqoj.onion	0.351		File Index's
7	4doqhu4gw5xoddmn.onion	0.296		Fresh Onions
8	directoryvi6plzm.onion	0.226		Tor Directory: A list of onion sites ordered by hostname
9	2xyqdwad2laqcd3v.onion	0.172		CB3ROB Tactical Data Services - TOR Darknet site listing
10	acjhxk5yqwnw2jdu.onion	0.167		CB3ROB Tactical Data Services - TOR Darknet site listing

図表 62. スナップショット B の出力次数の上位 10 ノード (n=53300)

rank	node	degree	centrality	title
1	56wr4dvq3abd2ivkf5z36nortvu7dgona55zqsihfaqo2aeg5er4moi	0.531	1717293	Fresh Onions
2	msydqstlz2kzerdg.onion	0.448	226046	Ahmia
3	zlal32teyptf4tvi.onion	0.407	793767	Fresh Onions
4	pejjyyh7rhv5ctyu.onion	0.387	005385	File Index's
5	2xyqdwad2laqcd3v.onion	0.172	948836	CB3ROB Tactical Data Services - TOR Darknet site listing
6	acjhxk5yqwnw2jdu.onion	0.171	860635	CB3ROB Tactical Data Services - TOR Darknet site listing
7	darkweb2zz7etehx.onion	0.128	013659	CB3ROB Tactical Data Services - TOR Darknet site listing
8	cb3robuo3hobodw6.onion	0.103	773054	CB3ROB Tactical Data Services
9	darkjmgasio2e3cq3.onion	0.102	666091	CB3ROB Tactical Data Services - TOR Darknet site listing
10	3bbaaaccczcbdddz.onion	0.102	253326	OnionLand Search

図表 63. スナップショット C の出力次数の上位 10 ノード (n=70684)

rank	node	degree	centrality	title
1	visiwnqyii4r5f5l.onion	0.354		VisiTOR
2	pejjyyh7rhv5ctyu.onion	0.280		File Index's
3	msydqstlz2kzerdg.onion	0.158		Ahmia
4	2xyqdwad2laqcd3v.onion	0.150		CB3ROB Tactical Data Services - TOR Darknet site listing
5	3bbad7fauom4d6sgppalyqddsqb5u5p56b5k5uk2zxsy3d6ey2joi	0.147		Privacy Policy Onionland Search Engine
6	acjhxk5yqwnw2jdu.onion	0.143		CB3ROB Tactical Data Services - TOR Darknet site listing
7	dhosting4xxoydyaivckq7tsmtgi4wfs3flpeyitekkmqwu4v4r46syd.	0.136		Daniel's Hosting
8	3bbaaaccczcbdddz.onion	0.116		OnionLand Search
9	cb3rob5vwac2dtyc.onion	0.112		CB3ROB Tactical Data Services - TOR CHAT
10	cb3robuo3hobodw6.onion	0.099		CB3ROB Tactical Data Services

図表 64. スナップショット D の出力次数の上位 10 ノード (n=103674)

rank	node	degree	centrality	title
1	msydqstlz2kzerdg.onion	0.245		Ahmia
2	visicgrfb443cqh.onion	0.236		Phishing
3	msyd6emf7clejhld.onion	0.212		Ahmia
4	pejjyyh7rhv5ctyu.onion	0.204		File Index's
5	42yn43ahvsm7tonn.onion	0.084		DarkLib register
6	dhosting4xxoydyaivckq7tsmtgi4wfs3flpeyitekkmqwu4v4r46syd.	0.079		Daniel's Hosting - Info
7	3bbaaaccczcbdddz.onion	0.074		OnionLand Search
8	3bbad7fauom4d6sgppalyqddsqb5u5p56b5k5uk2zxsy3d6ey2joi	0.066		Privacy Policy Onionland Search Engine
9	dhosting4okcs22v.onion	0.056		Daniel's Hosting
10	kj7fx3wribkd4ara.onion	0.055		n/a

図表 65. スナップショット E の出力次数の上位 10 ノード (n=54509)

rank	node	degree	centrality	title
1	msydqstlz2kzerdg.onion	0.476		Ahmia
2	msyd6emf7clejhld.onion	0.272		Ahmia
3	3bbaaaccczcbdddz.onion	0.093		OnionLand Search
4	jo7np7o7hhwksse.onion	0.043		Uncensored Hidden Wiki
5	visicgrfb443cqh.onion	0.030		Phishing
6	runionwe25lxr3is.onion	0.019		Runion
7	onioofr4na4kvarv.onion	0.018		n/a
8	wikilink77h7lrbj.onion	0.018		Warez Wiki Links
9	oniot2zvfcp4lpc.onion	0.017		n/a
10	h7xoxpbd3fjlru6.onion	0.016		n/a

図表 66. スナップショット F の出力次数の上位 10 ノード (n=17464)

rank	node	degree	centrality	title
1	freshonionleyrre.onion	0.058		Fresh Onions
2	r3sx53i7zj7b4a54.onion	0.052		Savage Rape ~ Candy Land ~ Secure Onions
3	zplubsflovgb2edv6tfu37frazttimtv6l4zee4al7exfdjw3frpgyd.oni	0.052		Young Child Pics Porno ~ Loliwood Studios ~ Fresh Onions
4	thsfpvk4aqu5hfbgjcwqroyziooygvkul3auh6jppakgmtb54hhmfac	0.052		Zoo Pussy : Ilovecp Onion : Onion Wiki
5	enrghdjlivscr2zmepvdw7t4cbnjbbmg7xxu55hdd5fq6zgtjitzoqd.c	0.052		Watch Porn Xxx - Boys Tower - Onions List
6	p7o4vcxp63hd2gi4.onion	0.052		Young Cum ~ Boy Forum ~ Onion
7	r6njocvduhcjxjqst3hdijugtyvg52hclataoa3a6rq7xf2cdzvad.on	0.052		Petite Young Pussy - Bd Company Gallery - Onion Pool
8	grojq2klhdwfmhmqkxhpdqkqoadtfdcn3ocg5mk64223srbcdcwjy	0.052		Candydolls Cp : Boy Vid : Listed Onions
9	ti47sqgww4dmdmgtt5jpmxcibwobokgd3brqncqnsqorcgotwcvbz6v	0.052		What Happened To Boyvid 4 0 Ilovecp Site Porn Onions
10	7d5f76i3tp5q46cvkmvp75yrtowmnb4avew5vmmf5d6f2usy27mii	0.052		Young And Old Porn ~ Bonanza Forum ~ Onion Verified

図表 67. スナップショット A-F の出力次数の上位 10 ノード (n=172740)

rank	node	degree	centrality	title
1	msydstlz2kzerdg.onion	0.387		Ahmia
2	56wr4dvq3abd2ivkf5z36nortvu7dgonas5zqsihfaqo2aeg5er4moi	0.188		Fresh Onions
3	msyd6emf7clejhld.onion	0.180		Ahmia
4	zlal32teyptf4tvi.onion	0.170		Fresh Onions
5	visiwnqyii4r5f5l.onion	0.145		VisiTOR
6	visicgxfb443cqh.onion	0.142		Phishing
7	3bbaaaccczcbdddz.onion	0.132		OnionLand Search
8	msydstj6vtexpg.onion	0.124		Ahmia
9	pejjyyh7rhv5ctyu.onion	0.123		File Index's
10	msydstlzbk3tr5q.onion	0.121		Ahmia

付録 B 3. PageRank による中心性指標一覧

図表 68. スナップショット A の PageRank の上位 10 ノード (n=48360)

rank node	page rank	title
1 zlal32teyptf4tvi.onion	0.045	Fresh Onions
2 56wr4dvq3abd2ivkf5z36nortvu7dgona55zqsihfaqo2aeg5er4moi	0.041	Fresh Onions
3 msydqstlz2kzerdg.onion	0.027	Ahmia
4 msydqstjd6vtexpq.onion	0.026	Ahmia
5 msydqstlzbk3tr5q.onion	0.026	Ahmia
6 directoryvi6plzm.onion	0.024	Tor Directory: A list of onion sites ordered by hostname
7 4doqhu4gw5xoddmn.onion	0.023	Fresh Onions
8 wrrkz262g55scqoj.onion	0.022	File Index's
9 zdxgqrvvwpnuj2n.onion	0.010	A list of onion sites ordered by title
10 3wfzef3ql23dngkh.onion	0.009	Dark Onions - The list of onion sites

図表 69. スナップショット B の PageRank の上位 10 ノード (n=53300)

rank node	page rank	title
1 zlal32teyptf4tvi.onion	0.071907517	Fresh Onions
2 56wr4dvq3abd2ivkf5z36nortvu7dgona55zqsihfaqo2aeg5er4moi	0.067608975	Fresh Onions
3 msydqstlz2kzerdg.onion	0.051822209	Ahmia
4 pejjyyh7rhv5ctyu.onion	0.046407093	File Index's
5 2xyqdwad2laqcd3v.onion	0.013712177	CB3ROB Tactical Data Services - TOR Darknet site listing
6 acjhxk5yqwnw2jdu.onion	0.013687437	CB3ROB Tactical Data Services - TOR Darknet site listing
7 dhosting4xxoydyaivckq7tsmtgi4wfs3flpeyitekkmqwu4v4r46syd.	0.013090016	Daniel's Hosting - Info
8 darkweb2zz7etehx.onion	0.009917522	CB3ROB Tactical Data Services - TOR Darknet site listing
9 cratedvnn5z57xhl.onion	0.009711358	The onion crate - Tor hidden service indexThe onion crate - Tor hidden service index
10 msyd5324jqjq4jw6.onion	0.009117583	Ahmia

図表 70. スナップショット C の PageRank の上位 10 ノード (n=70684)

rank node	page rank	title
1 visiwngyii4r5f5l.onion	0.091	VisiTOR
2 pejjyyh7rhv5ctyu.onion	0.067	File Index's
3 dhosting4xxoydyaivckq7tsmtgi4wfs3flpeyitekkmqwu4v4r46syd.	0.017	Daniel's Hosting - Info
4 msydqstlz2kzerdg.onion	0.017	Ahmia
5 msyd5324jqj4jw6.onion	0.014	Ahmia
6 3bbad7fauom4d6sgppalyqddsqb5u5p56b5k5uk2zxsy3d6ey2jot	0.014	Onionland Search Engine
7 2xyqdwad2laqcd3v.onion	0.013	CB3ROB Tactical Data Services - TOR Darknet site listing
8 jld3zkuo4b5mbios.onion	0.012	All Onion Services
9 acjhxk5yqwnw2jdu.onion	0.012	CB3ROB Tactical Data Services - TOR Darknet site listing
10 3bbaaaccczcbdddz.onion	0.011	OnionLand Search

図表 71. スナップショット D の PageRank の上位 10 ノード (n=103674)

rank node	page rank	title
1 visicgxrfb443cqh.onion	0.060	Phishing
2 pejjyyh7rhv5ctyu.onion	0.048	File Index's
3 msydqstlz2kzerdg.onion	0.029	Ahmia
4 msyd6emf7clejhld.onion	0.027	Ahmia
5 42yn43ahvsm7tonn.onion	0.015	DarkLib register
6 kj7fx3wribkd4ara.onion	0.011	n/a
7 5uqoveemnt3nlb3o.onion	0.011	n/a
8 dhosting4xxoydyaivckq7tsmtgi4wfs3flpeyitekkmqwu4v4r46syd.	0.009	Daniel's Hosting - Info
9 3bbaaaccczcbdddz.onion	0.007	OnionLand Search
10 jo7np7o7hhwksse.onion	0.007	Uncensored Hidden Wiki

図表 72. スナップショット E の PageRank の上位 10 ノード (n=54509)

rank node	page rank	title
1 msydqstlz2kzerdg.onion	0.104	Ahmia
2 msyd6emf7clejhld.onion	0.054	Ahmia
3 3bbaaaccczcbdddz.onion	0.015	OnionLand Search
4 jo7np7o7hhwksse.onion	0.011	Uncensored Hidden Wiki
5 visicgxrfb443cqh.onion	0.007	Phishing
6 oa7afprjhd6ffocc.onion	0.003	n/a
7 runionwe25lxr3is.onion	0.002	Runion
8 il2tacbrarmadm4q.onion	0.002	Wiki Link Hidden Tor .onion urls directories
9 m6kshkn2vcovss35.onion	0.002	Reaper Mafia
10 2f7f4rlcu62w4zph.onion	0.002	Tor66 - Support Us by donation

図表 73. スナップショット F の PageRank の上位 10 ノード (n=17464)

rank node	page rank	title
1 freshonionleyrre.onion	0.013	Fresh Onions
2 runionwe25l3is.onion	0.011	Runion
3 wikiheyd6zuc54j.onion	0.010	Hidden Wiki Deep Web Links / Dark Web Links
4 dwlonion3o3pjql.onion	0.009	Darkode Reborn Dark Web Link Deep web Onion Links Darknet News
5 msydqstlz2kzerdg.onion	0.008	Ahmia
6 c4z3uzmrifiszewo.onion	0.008	Onion Dir List
7 onionl5znxfreym4.onion	0.007	Onion Dir
8 onion2pwcaijosix.onion	0.007	Onion Dir
9 hdwiy54noz34b3tf.onion	0.007	HD Wiki
10 dark2vyfyvndt4xc.onion	0.005	What is bitcoin? - Dark Matter Project

図表 74. スナップショット A-F の PageRank の上位 10 ノード (n=172740)

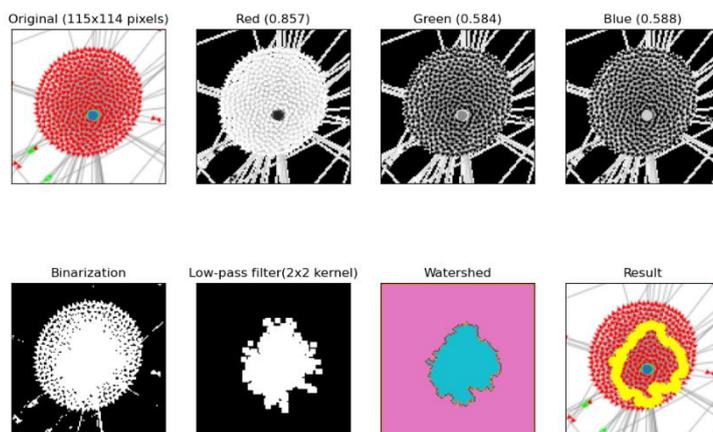
rank node	page rank	title
1 msydqstlz2kzerdg.onion	0.042	Ahmia
2 msyd6emf7clejhld.onion	0.022	Ahmia
3 zlal32teyptf4tvi.onion	0.014	Fresh Onions
4 56wr4dvq3abd2ivkf5z36nortvu7dgona55zsqsihfaqo2aeg5er4moi	0.014	Fresh Onions
5 visiwngy4r5f5l.onion	0.013	VisiTOR
6 visicgrfb443cqh.onion	0.013	Phishing
7 dhosting4xxoydyavckq7tsmtgi4wfs3flpeyitekkmqwu4v4r46syd.	0.012	Daniel's Hosting
8 3bbaaacczcbdddz.onion	0.011	OnionLand Search
9 3bbad7fauom4d6sgppalyqddsqb5u5p56b5k5uk2zxsy3d6ey2jot	0.009	Onionland Search Engine
10 42yn43ahvsm7tonn.onion	0.009	DarkLib

付録 C

本稿で使用した各スナップショットから目視で抽出した標本の RGB 各成分の分解と、Watershed 法により自動認識された領域、そして視覚化されたダークウェブハイパーリンクネットワークグラフの拡大図をそれぞれ示す。

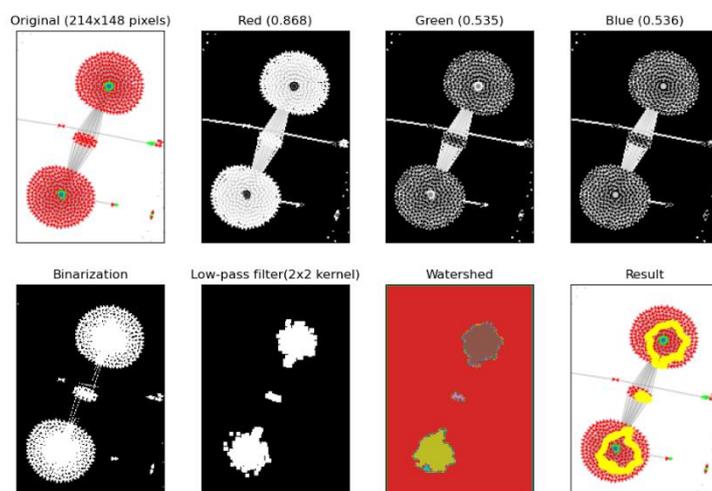
領域の自動認識には、各標本を二値化した上で 2×2 カーネルのローパスフィルタによりノイズ成分を除去し、20%まで領域を収縮してから、各マーカー領域に Watershed 法を適用して領域の自動判定を行った。目視で抽出した標本が自動認識の場合どのように判別されるかを確認できる。

付録 C 1. 基準標本群とした標本一覧



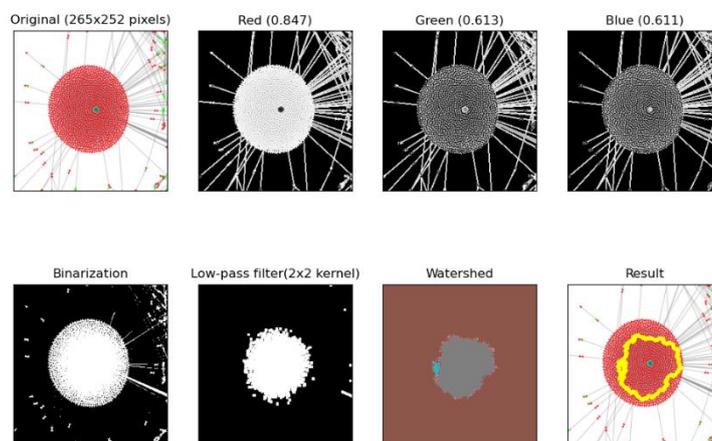
図表 75. 標本 A の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



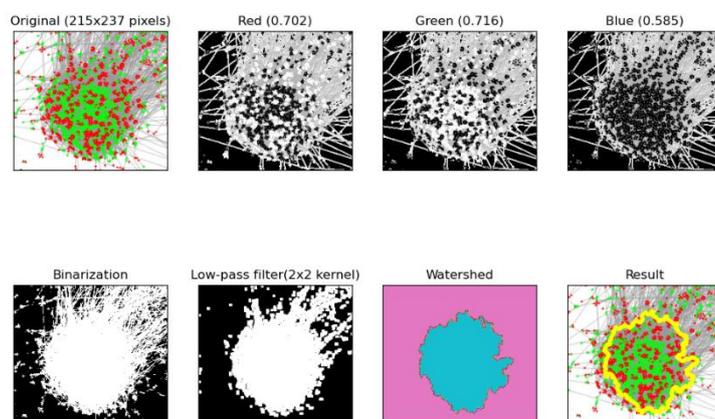
図表 76. 標本 B の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



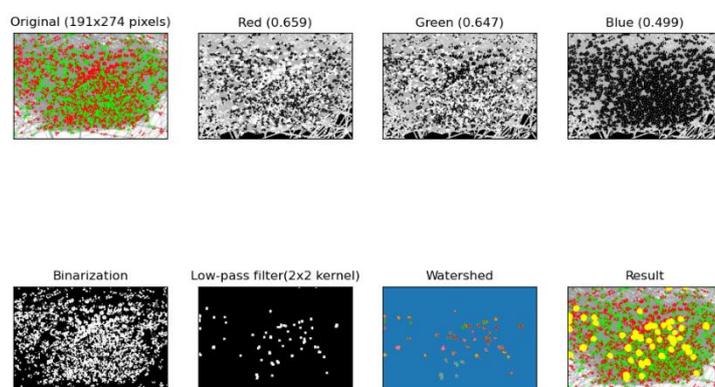
図表 77. 標本 C の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



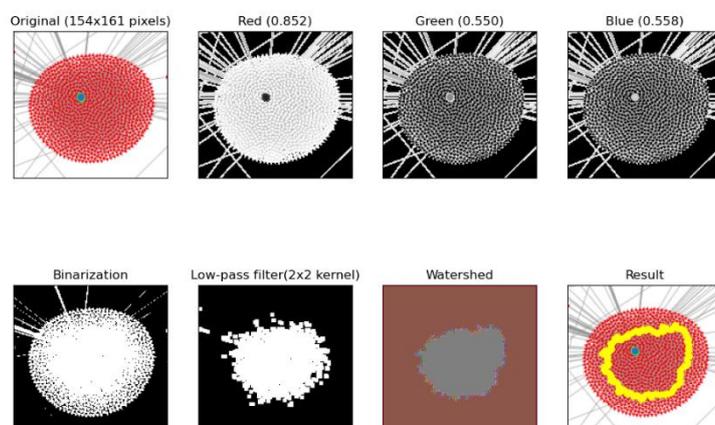
図表 78. 標本 D の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



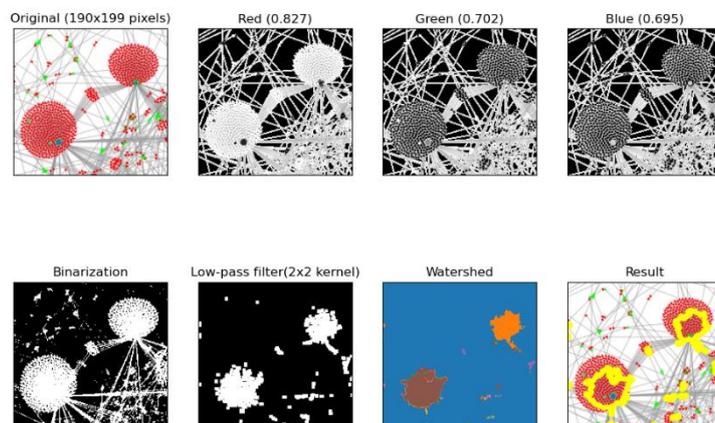
図表 79. 標本 E の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



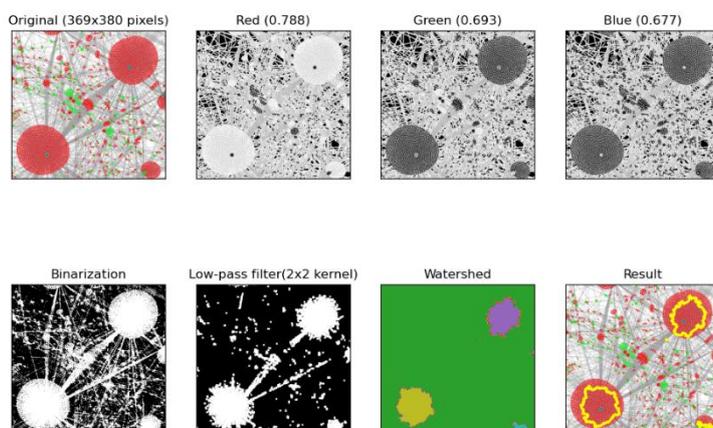
図表 80. 標本 F の色成分

RGB 各成分の分解と，Watershed 法により自動認識された領域



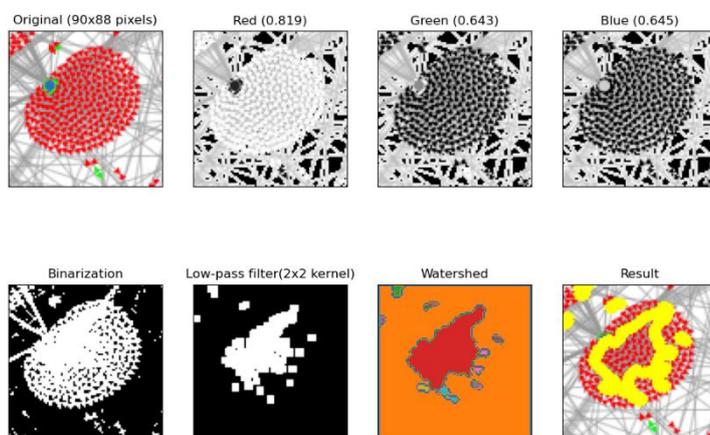
図表 81. 標本 G の色成分

RGB 各成分の分解と，Watershed 法により自動認識された領域



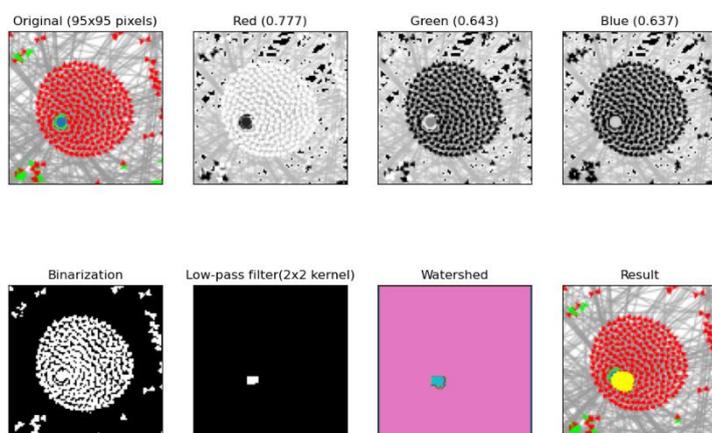
図表 82. 標本 H の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



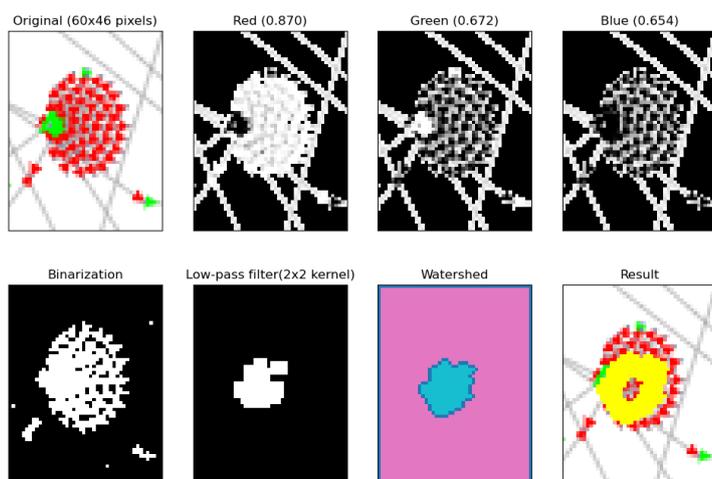
図表 83. 標本 I の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



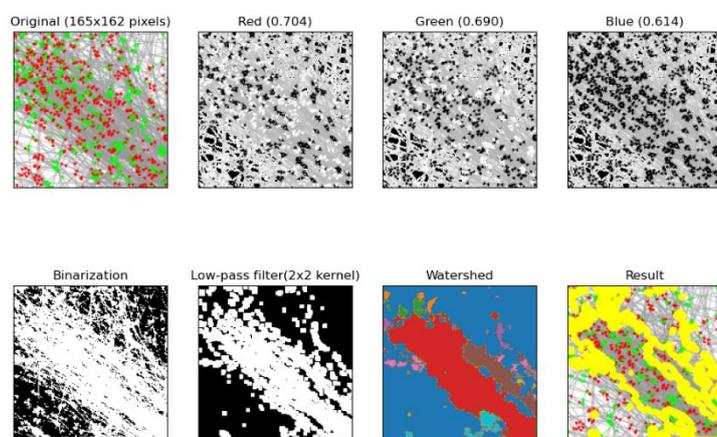
図表 84. 標本 J の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



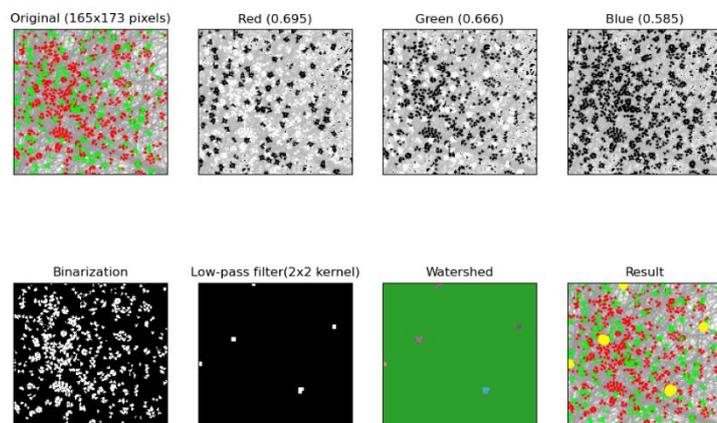
図表 85. 標本 K の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



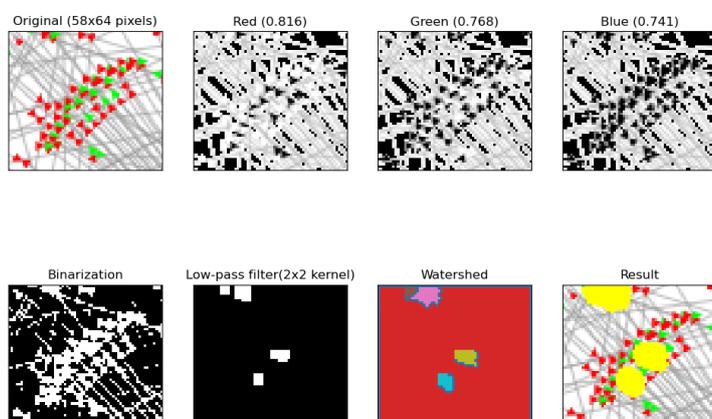
図表 86. 標本 L の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



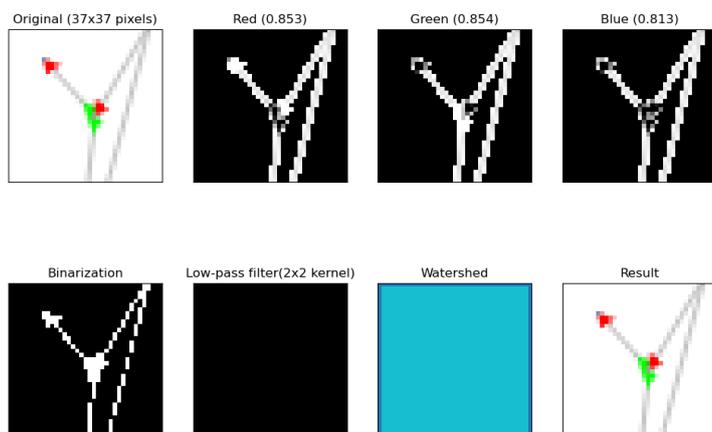
図表 87. 標本 M の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



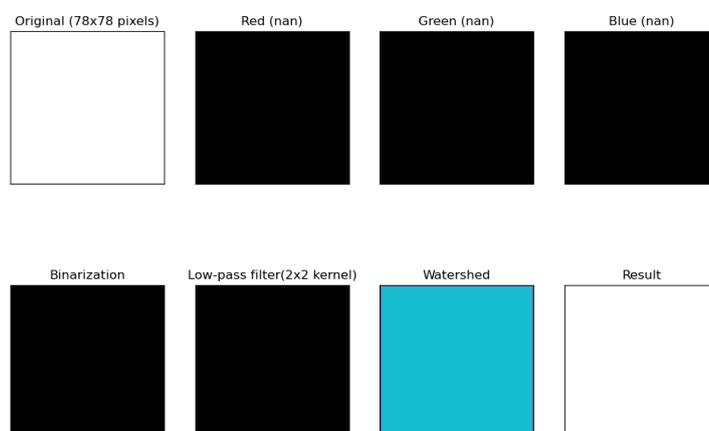
図表 88. 標本 N の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



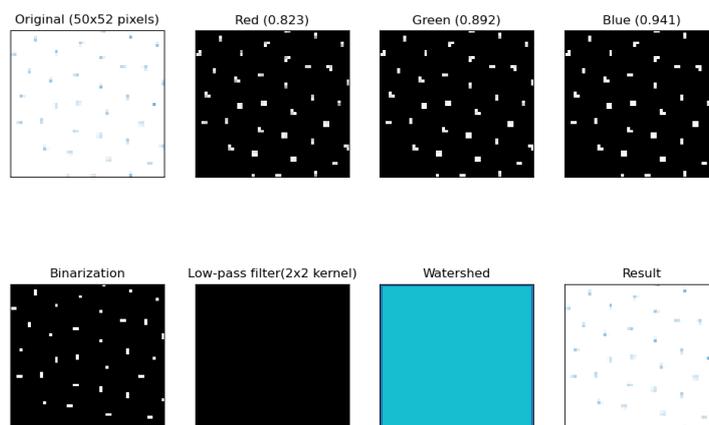
図表 89. 標本 O の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



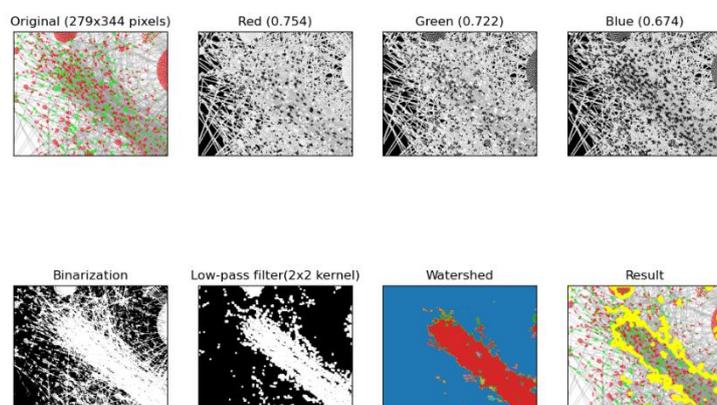
図表 90. 標本 P の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



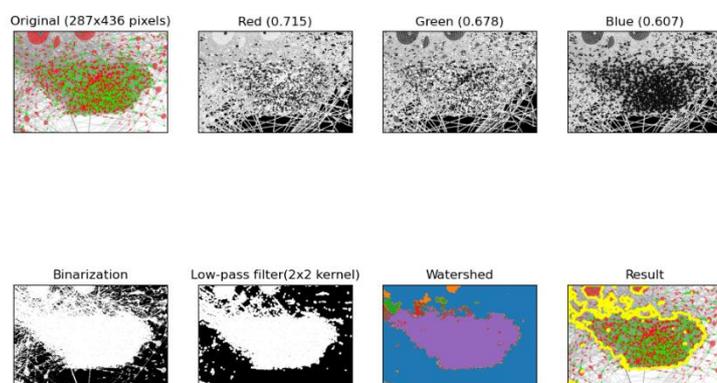
図表 91. 標本 Q の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



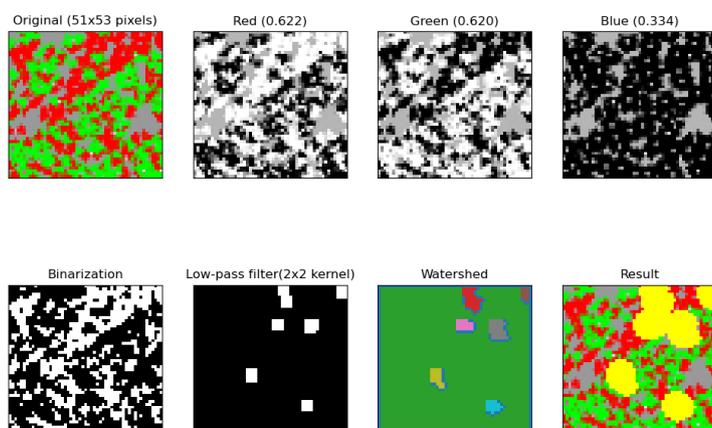
図表 92. 標本 R の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



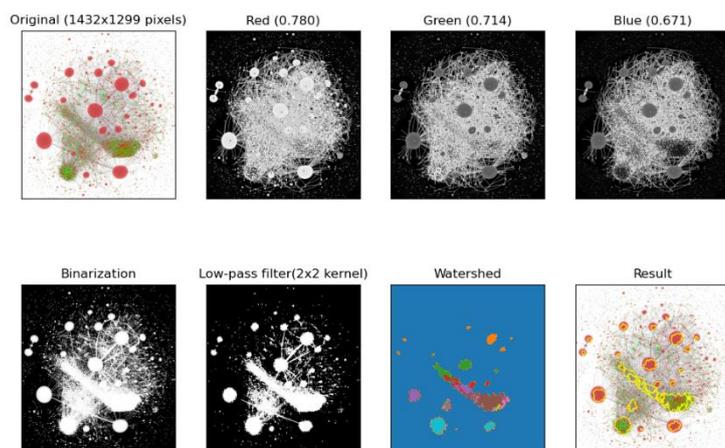
図表 93. 標本 S の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



図表 94. 標本 T の色成分

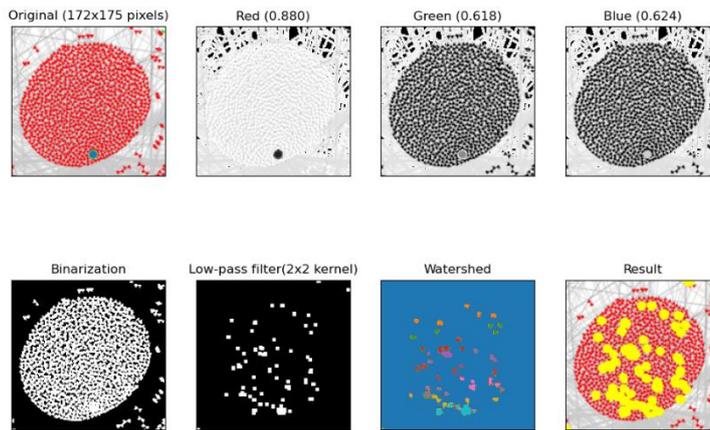
RGB 各成分の分解と、Watershed 法により自動認識された領域



図表 95. 標本 U の色成分

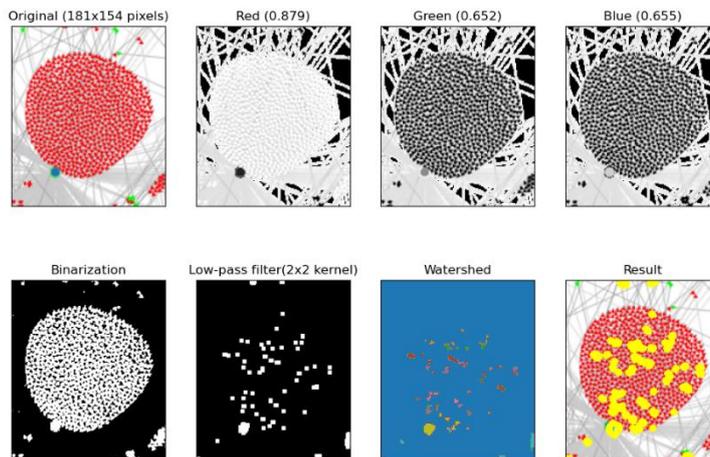
RGB 各成分の分解と、Watershed 法により自動認識された領域

付録 C 2. スナップショット A の標本一覧



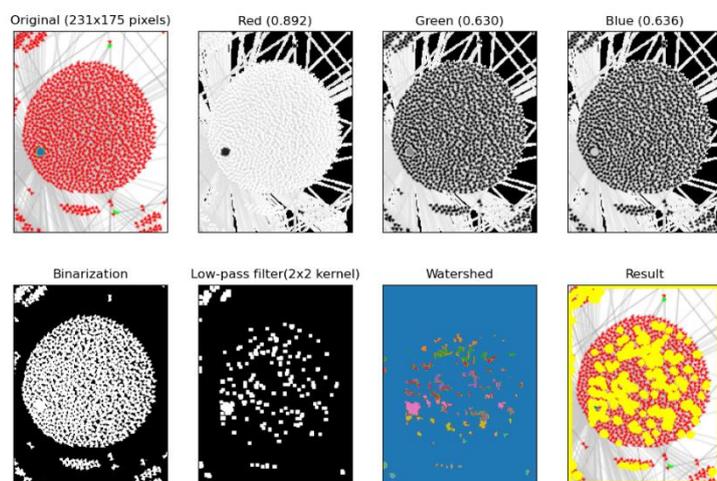
図表 96. 標本 a0 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



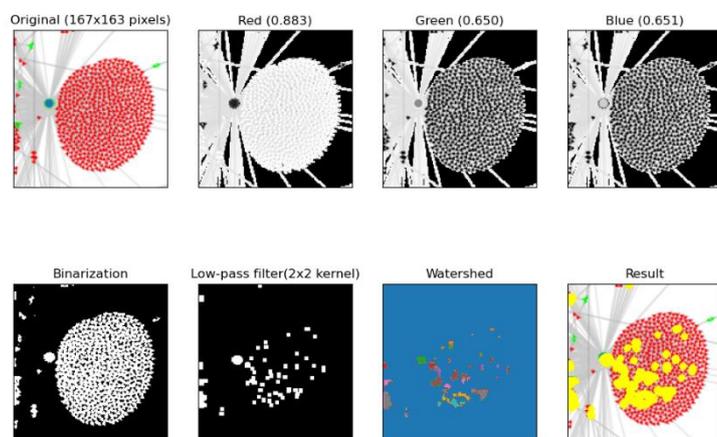
図表 97. 標本 a1 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



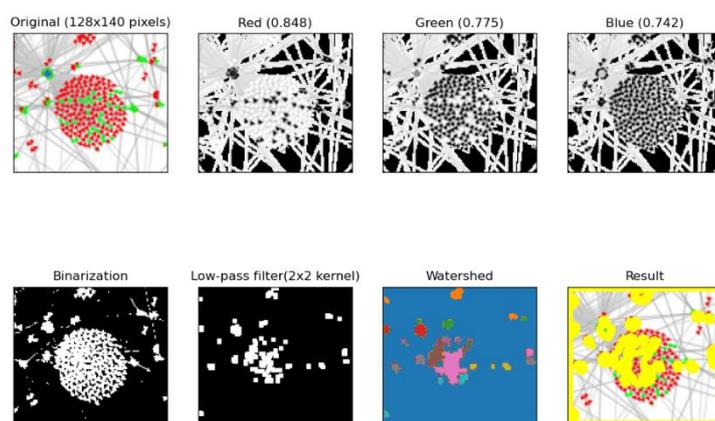
図表 98. 標本 a2 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



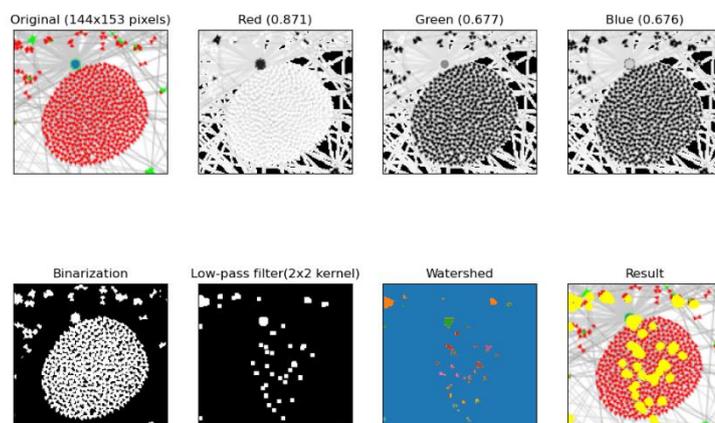
図表 99. 標本 a3 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



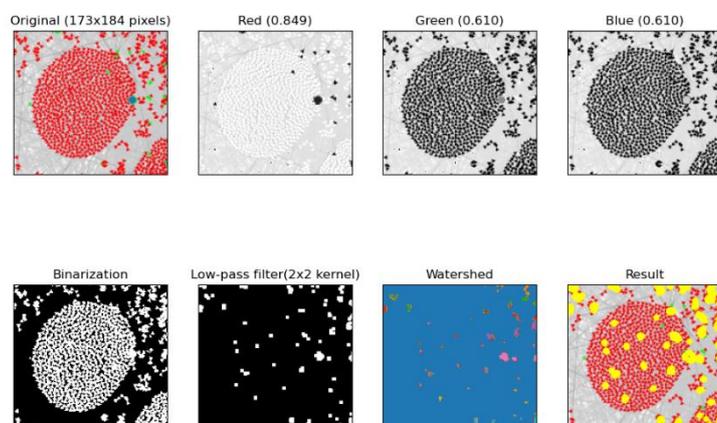
図表 100. 標本 a4 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



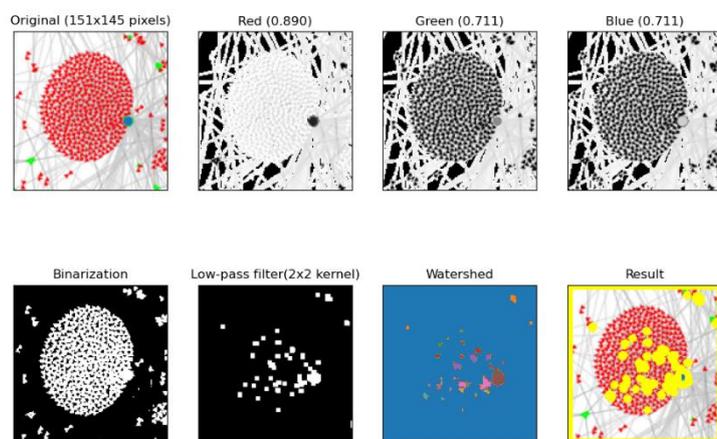
図表 101. 標本 a5 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



図表 102. 標本 a6 の色成分

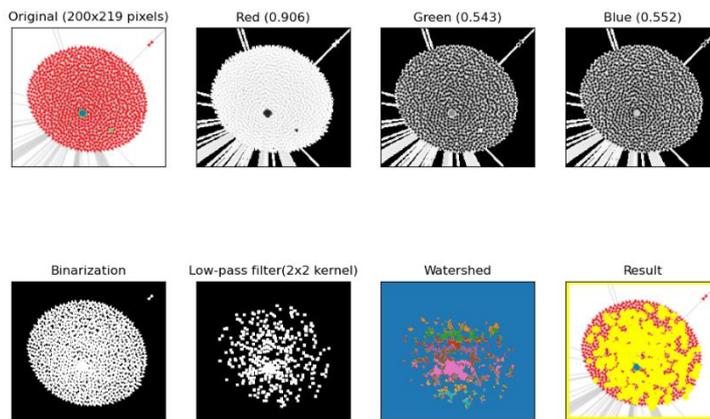
RGB 各成分の分解と、Watershed 法により自動認識された領域



図表 103. 標本 a7 の色成分

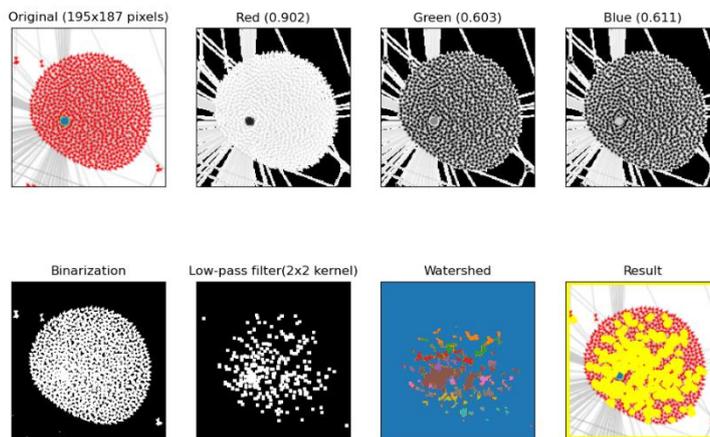
RGB 各成分の分解と、Watershed 法により自動認識された領域

付録 C 3. スナップショット B の標本一覧



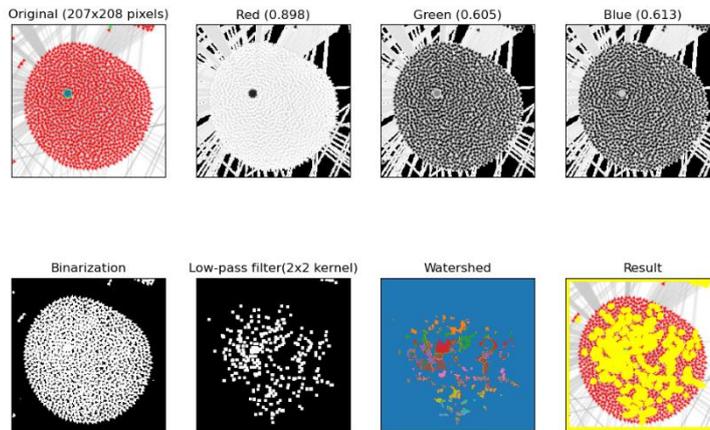
図表 104. 標本 b0 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



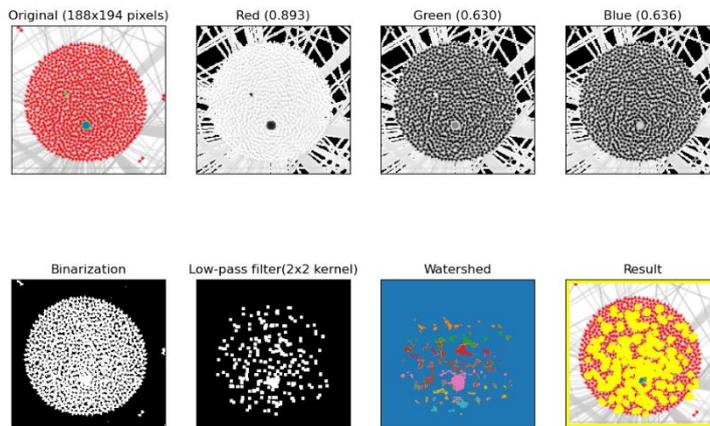
図表 105. 標本 b1 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



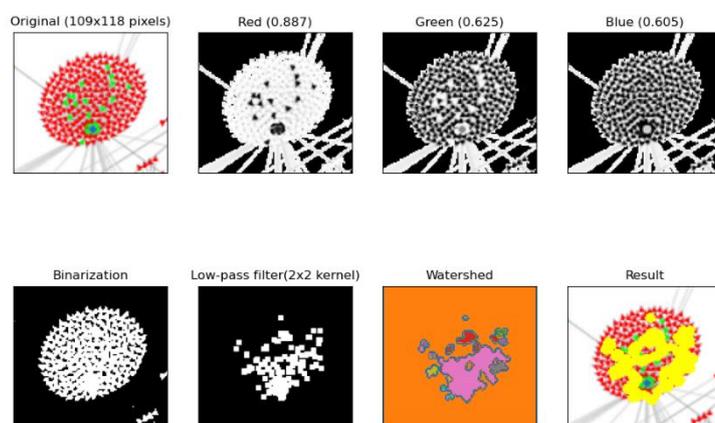
図表 106. 標本 b2 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



図表 107. 標本 b3 の色成分

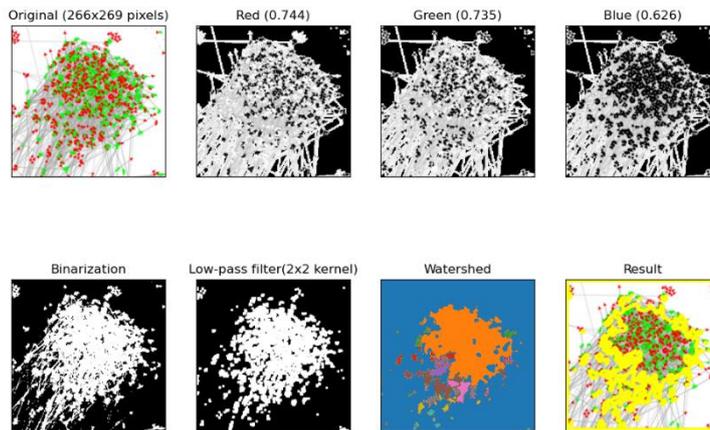
RGB 各成分の分解と、Watershed 法により自動認識された領域



図表 108. 標本 b4 の色成分

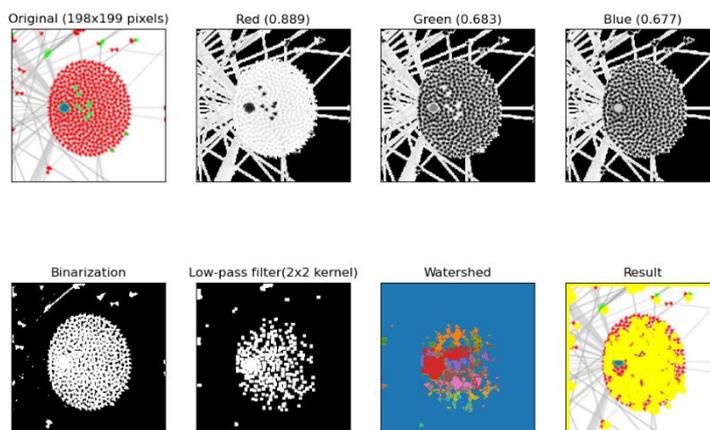
RGB 各成分の分解と, Watershed 法により自動認識された領域

付録 C 4. スナップショット C の標本一覧



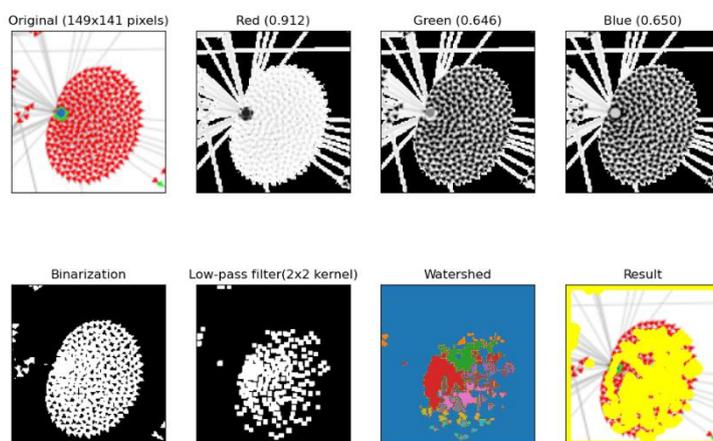
図表 109. 標本 c0 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



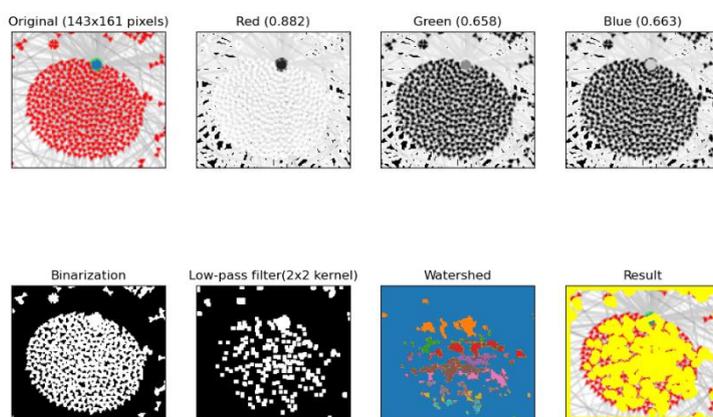
図表 110. 標本 c1 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



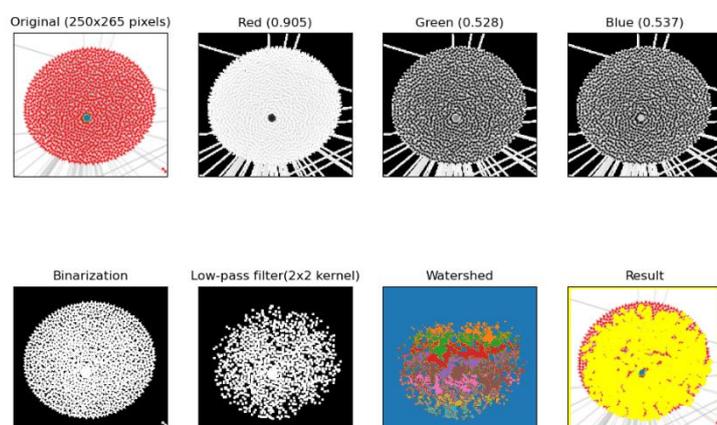
図表 111. 標本 c2 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



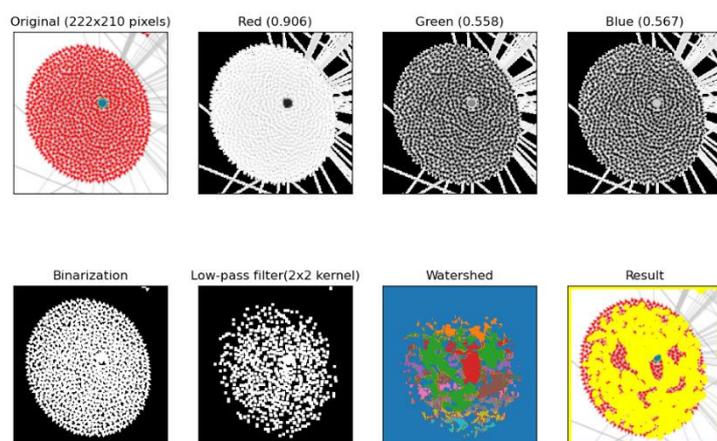
図表 112. 標本 c3 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



図表 113. 標本 c4 の色成分

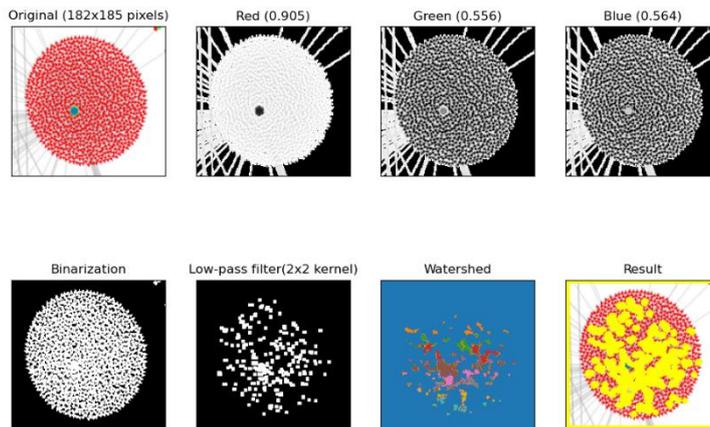
RGB 各成分の分解と、Watershed 法により自動認識された領域



図表 114. 標本 c5 の色成分

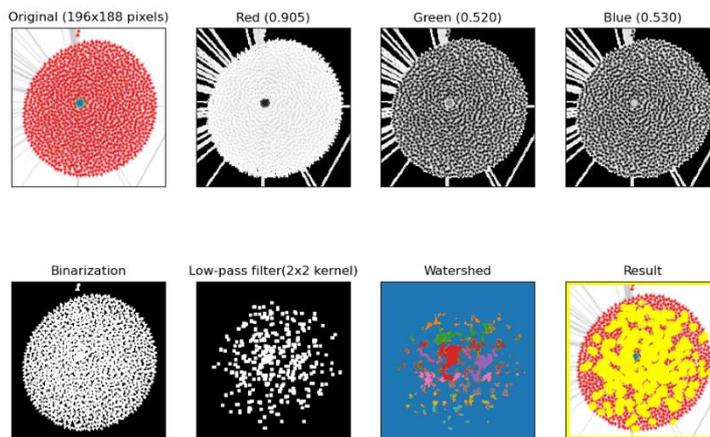
RGB 各成分の分解と、Watershed 法により自動認識された領域

付録 C 5. スナップショット D の標本一覧



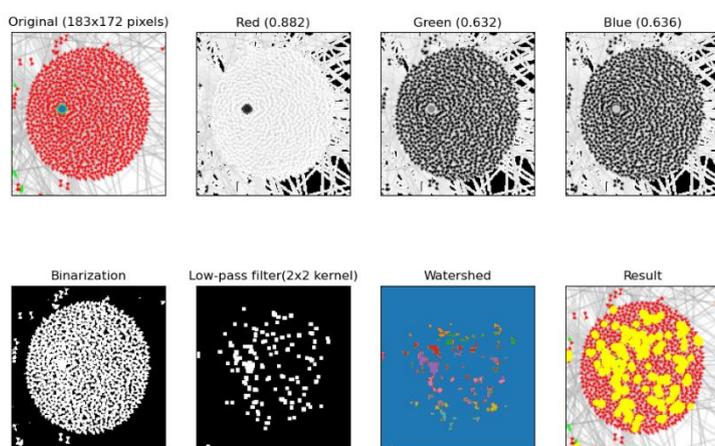
図表 115. 標本 d0 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



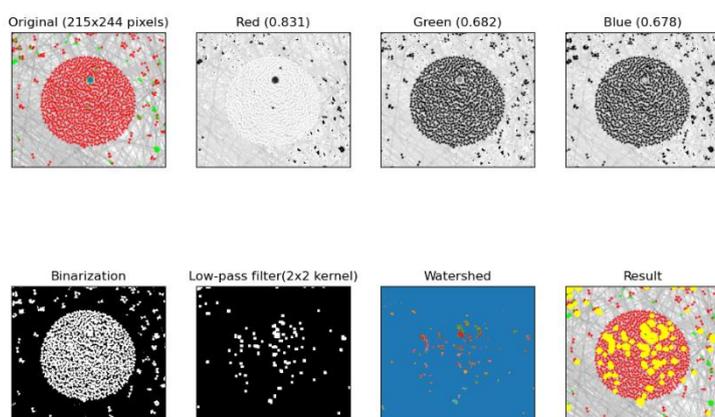
図表 116. 標本 d1 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



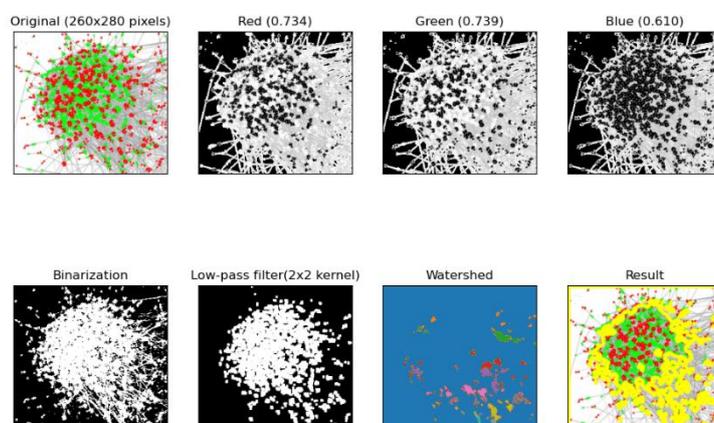
図表 117. 標本 d2 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



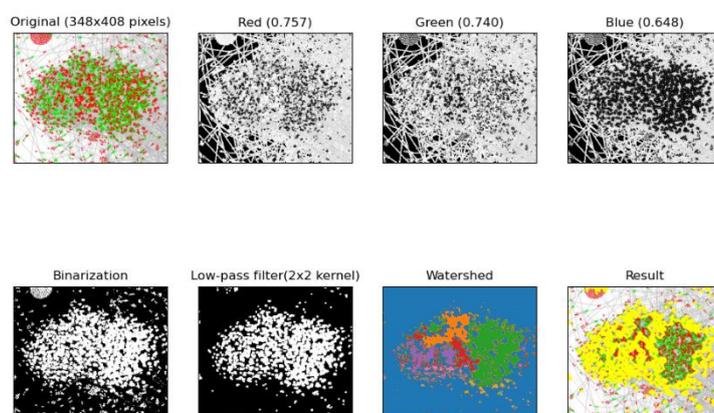
図表 118. 標本 d3 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



図表 119. 標本 d4 の色成分

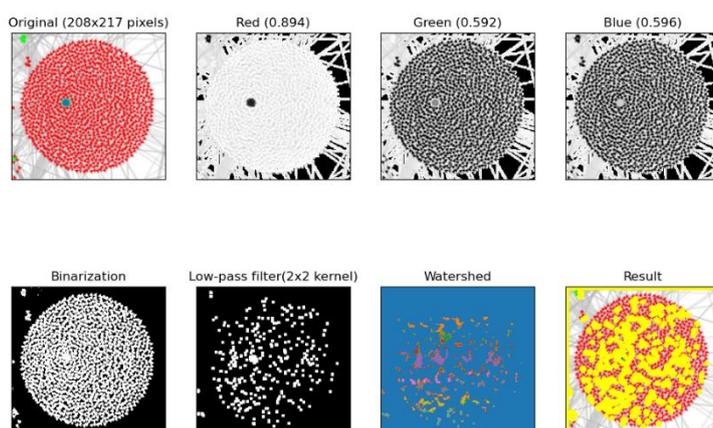
RGB 各成分の分解と、Watershed 法により自動認識された領域



図表 120. 標本 d5 の色成分

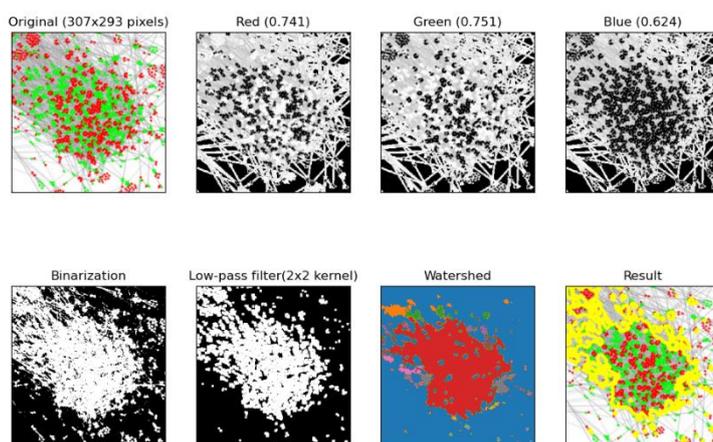
RGB 各成分の分解と、Watershed 法により自動認識された領域

付録 C 6. スナップショット E の標本一覧



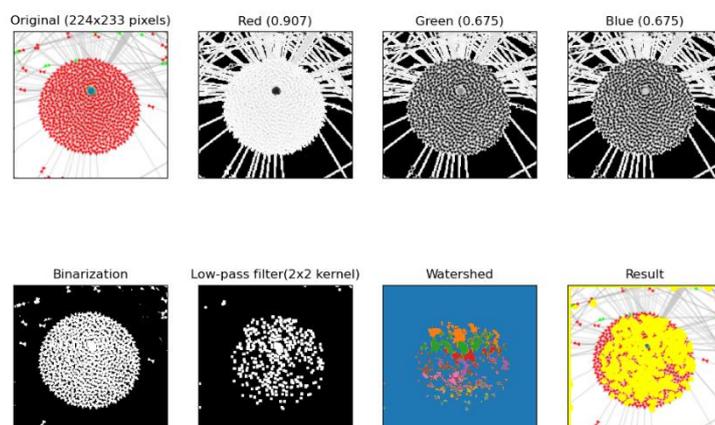
図表 121. 標本 e0 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



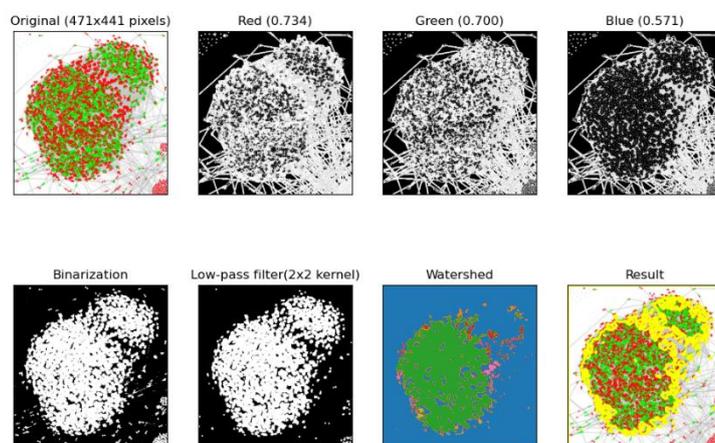
図表 122. 標本 e1 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



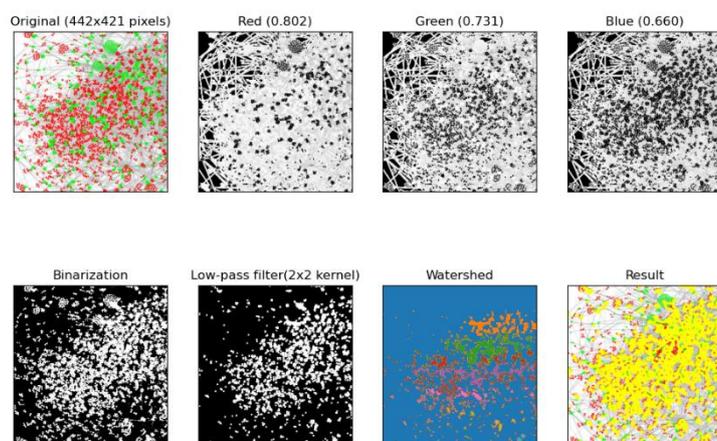
図表 123. 標本 e2 の色成分

RGB 各成分の分解と、Watershed 法により自動認識された領域



図表 124. 標本 e3 の色成分

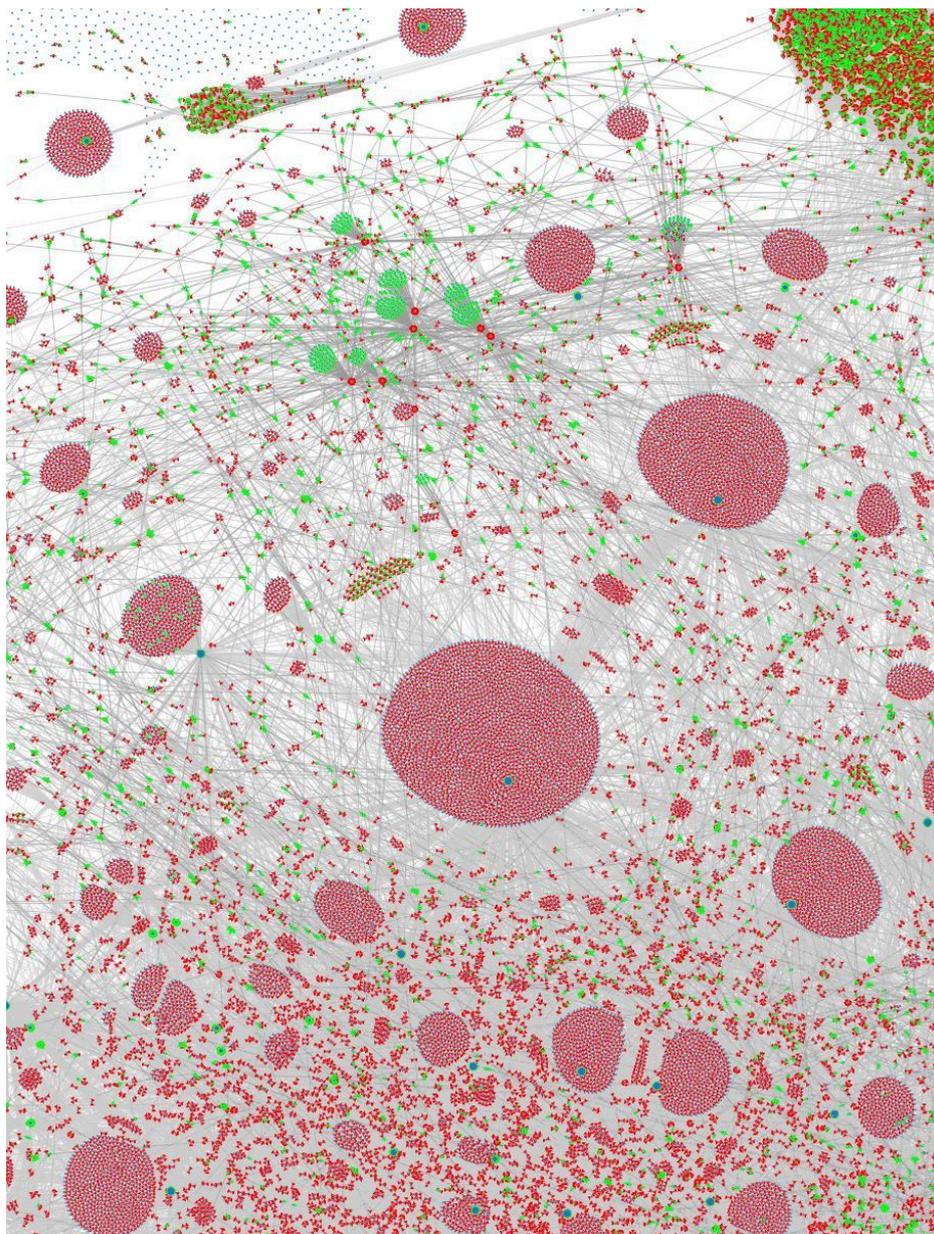
RGB 各成分の分解と、Watershed 法により自動認識された領域



図表 125. 標本 e4 の色成分

RGB 各成分の分解と, Watershed 法により自動認識された領域

付録C 7. ダークウェブハイパーリンクネットワーク グラフの拡大図



図表 126. ダークウェブハイパーリンクネットワークグラフの拡大図
期間は 2018 年 6 月 1 日から 2021 年 1 月 30 日まで