Vol.2018-IOT-43 No.10
Vol.2018-SPT-30 No.10
2018/9/28

Dark Webのコンテンツ分析とつながりの解明

杉生 貴成[1]　猪俣 敦夫[1]

概要　近年，Tor 等の匿名化ネットワークを使用して送信元を隠蔽して接続する手段を用いて接続，また検索エンジンにインデックスされないダークウェブ（Dark Web）と呼ばれる Web サイト群が注目されている．Dark Web は麻薬市場，銃器関連の密売，児童ポルノ，脆弱性情報，Booter（DDoS 請負サービス）のような違法行為やサービスのための理想的なホスティングの場となっている．世界中の研究者や法執行機関，セキュリティ会社，学術機関で調査を行っているが，効率的な調査方法が確立されておらず，調査に時間がかかる問題がある．本研究では Dark Web 上のサイトの一つである Ichidan と呼ばれている Web サイトを利用し大量の Onion ドメイン名を取得し分析を行った．取得した Onion ドメインに全てに対してスクレイピングを行い，トップページをダウンロードした．ダウンロードしたトップページのテキストから Onion ドメインを 6 つのカテゴリに大別した．そしてハイパーリンクのつながりから Onion ドメインをノードとし被ハイパーリンクの方向を矢印とした有向グラフを作成し Dark Web 同士の接続状況が容易に捉えられるようにした．そのグラフに大別した 6 つのカテゴリを合わせることでコンテンツごとのつながりや特性について解明しようと試みた．

Dark Web content analysis and probe connection

TAKAAKI SUGIU[1]　ATSUO INOMATA[1]

1.　はじめに

インターネットの世界において検索エンジンなどで収集できる情報は Surface Web(表層 Web)と呼ばれており，ブログやニュースサイト・SNS がこれにあたる．Surface Web は，外部からのハイパーリンクによるつながりや自ら申請することで意図的に検索エンジンにインデックスされる．それに対して検索エンジンでは見つけることのできない情報を Deep Web(深層 Web)と呼ばれている．例えば，Web サイトの設定によりクローリングしないまたは,パスワード設定がされている等によって検索エンジンが辿り着けない Web サイトを指す．カルフォルニア大学の調査によると Deep Web は Surface Web に対して 400～550 倍の公開情報があるとの結果[1]を示している

Deep Web の領域内に Dark Web と呼ばれるインターネット空間がある．Dark Web は検索エンジンでは辿り着くことができないうえ，Tor(The Onion Router)や I2P(The Invisible Internet Project)のような匿名化ネットワークとそれに対応した専用のアプリケーションを使用して閲覧することができる[2]．例えば，Tor のような匿名化ネットワークは，複数のリレーエージェントを経由し，経由したサーバにログが残らない．また，アクセス経路上，出口部分以外のすべてが暗号化される．さらに追跡を難しくするため，Tor では一定時間ごとに経路も変更される[3]．よって非常に匿名性の高いネットワークである．そして活発な取引がある Dark Web の違法サイトでは，薬物や児童ポルノ，武器，クレジットカード情報などの売買が横行しサイバー攻撃などを請け負うサイトも存在する[4]．

1.東京電機大学　〒120-8551　東京都足立区千住旭町 5 番
E-mail: sugiu@isl.im.dendai.ac.jp, inomata@mail.dendai.ac.jp

2.　Tor について

2.1　Tor とは

Tor は米海軍調査研究所の出資によって考案されたオニオンルーティングによる匿名化ネットワークのことである．クライアントが一対一の関係で通信するような形態のネットワークである P2P 技術と，TCP／IP の通信を中継する SOCS 技術を組み合わせて実装されている．Tor のリレーノードは世界中にあり現在 6000[5]以上のサーバがある．そしてリレーノードはログを残さないうえ，一定時間で接続するノードを変更しさらに出口以外の全ての通信が暗号化されているため，通信元が不明になり非常に匿名性が高い．Tor のアクセス経路の概略を図 1 に示す．

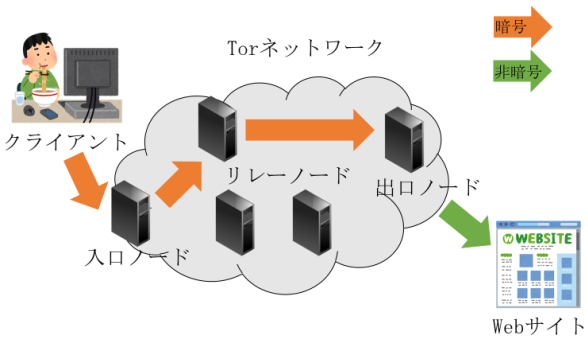

図 1　Tor のアクセス経路例

2.2　倫理ガイドライン

Tor Project には倫理ガイドライン[6]が存在しいくつかの行為が禁止されている．容認できない活動の例を以下に紹介する．

1. HSDir を実行し Onion のアドレスを収集し，それらに対して接続すること．なお，HSDir（hidden service directory）とはユーザが秘匿サービスを訪問するため

Vol.2018-IOT-43 No.10
Vol.2018-SPT-30 No.10
2018/9/28

のネットワークリレーである．[7].

2. 出口リレーノードを盗聴目的で設置し，トラフィックを改ざんすること．
3. 特定のサイトへの接続を終了すること．

これらのガイドラインを違反した研究チームは Tor のリレーから永久に追放されてしまった[8]．彼らは悪意のある匿名サービスの分類に関する研究をしており併せて Onion のアドレスを収集していた．

本研究では HSDir を使用せずまた，出口ノードも設置していないため Tor Project の倫理ガイドラインには違反していない．具体的に本研究では大量の Onion ドメインを収集したのではなく，Ichidan[9]を使用した分析を行っており，倫理ガイドラインに違反していないことをあらかじめ主張する．Ichidan とは Dark Web 用の Shodan[10]に類似した検索エンジンである．Shodan とはポートスキャンとバナーの調査により，インターネットに接続されている機器情報を収集して公開しているウェブサイトである[11]

3. Dark Web 関連の事件

日本において Dark Web に関連した事件を紹介する．コインチェック事件である．事件の概要は，2018 年 1 月下旬にコインチェックが不正アクセスを受け約 580 億円相当の仮想通貨 NEM（ネム）が流出した．その後，犯人とされる人物が Dark Web において NEM を別の仮想通貨に交換する資金洗浄を目的とした仮想通貨取引サイト設置し，流出した仮想通貨が全て取引されてしまったのである．犯人と思われる人物が開設した Web サイトを図 2 に示す．

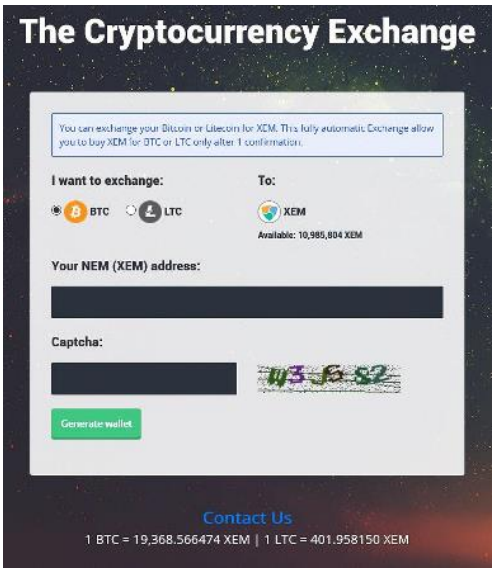

図 2　犯人と思われる人物が設置した Web サイト

次に，海外において Dark Web に関連した事件を紹介する．セキュリティ企業である 4Qi の研究者は Dark Web を調査し，41 ギガバイトの容量を持つファイルが Dark Web で発見され，その中に 14 億件分のユーザーネームとパスワードのセットが含まれていたと報告されている[12]．これらはパスワードなどが平文で保存されており，悪意のあるものの手に渡れば簡単に悪用ができる等の問題が考えられる．

4. 関連研究

Dark Web 上の不法なコンテンツを解明する研究として Ahmed T. Zulkarnine らの「Surfacing collaborated networks in dark web to find illicit and criminal content」が挙げられる[13]．彼らは，事前に定義したキーワードに基づいてインターネット上のウェブサイトを自動的にクローリングするウェブクローラーを開発した．それを Dark Web に応用し Dark Crawler というシステムを開発した．Dark Crawler を用いて Dark Web において過激派やテロリストの Web サイトや不法な Web サイトを発見し，どのように相互に接続されるかの初期発見を示した．Dark Web をネットワークグラフにマッピングした際，入次数・出次数のスコアの比較，人気キーワードに基づいて Dark Web の中心で人気のあるウェブサイトを発見した．人気のあるサイトを手動で調査するとロシアの麻薬市場や，Tor ネットワークで Dark Web に接続するためのドキュメントを提供している Web サイトであった．またランキングの上位では反アメリカ政府の立場と思われる主張をする政治的なブログもあった．本研究との違いは，対照する Web サイトの分類である．関連研究では過激派の Web サイトや反米的な思想を有するグループの監視・早期発見だが，本研究では Dark Web に存在する Web サイトを大まかに 6 つに分類する．ネットワークグラフを用いた可視化方法は関連研究と同じ手法だが，本研究ではコンテンツごとのつながりをユーザがより容易に把握するためノードにカテゴリ毎に色を定義してあるため，よりコンテンツ同士の接続状況を明確に提示できるようにした．

Dark Web を解析する分析するシステムを提案した論文として小野 諒人らの「HSDir の Snooping と秘匿サービスへのスキャンを組み合わせたダークウェブ分析システム」が挙げられる[14]. 彼らは Onion のアドレスを集めるため HSDir の Snooping を利用した．ユーザが匿名サービスを利用して Onion アドレスへ接続するが，その際 HSDir へアクセスし Introduction point などの情報を得る必要がある．この仕組を利用して観測用の HSDir を設置することでクライアントが実際にアクセスする Onion アドレスを収集することが可能になる．収集した Onion アドレスに対してサービススキャンを行うことで Dark Web をホスティングしているサーバの情報を収集することができる．スキャンの結果，不適切な状態でサービスを公開しているサーバが多く存在することが判明した．サービスの匿

名性で Onion アドレスを 3 種類に分類した際，最も匿名性が高いグループが最も多かった．また，クエリ数の多い上位 15 の Onion アドレスを手動でアクセスしコンテンツを確認した．それらのサイトは違法ドラッグを扱うブラックマーケットであった．本研究との違いは，Tor ネットワークにハニーポットのような観測機器を設置しておらず全て公開データのみで行い，また Dark Web に対してスキャン行為も行っていない．関連研究では違法サイトの発見方法としてアクセス数の頻度を使用しているが，本研究では予め定義したキーワードによる頻度に基づき判断している．

5. Dark Web 分析

本章では Dark Web を分析する目的と提案手法を説明する．本研究では Dark Web 上に存在するとされる違法サイトに加え，Dark Web を閲覧しているときによく見かけた Web サイトなど違法ではないサイトも含めた大まかにカテゴリを分類することである．そして分類したカテゴリ同士がどのように接続されているか，被リンクの解析による未知の Tor ドメインの発見，発見した Tor ドメインに実際にアクセスすることなく Web サイトのカテゴリを分類する．Dark Web のコンテンツを分析する手順を以下に示す．その後一つ一つ詳細に説明し結果を示す．

(1) Ichidan を用いて「http」が有効な Onion ドメイン名を取得
(2) 取得した Onion ドメイン全てに対してスクレイピングをしてトップページをダウンロード
(3) ページ内のテキストを用いて予め大別した 6 つのカテゴリに分類
(4) ページの内のハイパーリンクから有向グラフを作成
(5) 作成した有向グラフにカテゴリ分類結果を反映

5.1

（1） Ichidan を用いて「http」が有効な Onion ドメイン名を取得

検索方法はクエリ部分にプロトコル名等を付与する．今回 HTTP が有効な Onion ドメインを取得するため「http://ichidanv34wrx7m7.onion&query=http」となる．すると HTTP のサービスが有効な Onion ドメインの一覧が列挙される． なお，現在 Ichidan は閉鎖さ，アクセスできない状態にある．

（2）取得した Onion ドメイン全てに対してスクレイピングをしてトップページのダウンロード.

Python のライブラリである urllib を用いてスクレイピングする．Tor のプログラムを実行すると localhost の 9050 番ポートが Listen 状態になり，プロキシの設定をすることで Dark Web にプログラムから接続できるようになる．今回取得したデータの概要を表 1 に示す.

表 1　取得したデータの概要

取得した総 Web サイト数	8,291
リーチブルな Web サイト数	5,375
収集期間	2018/01/30~2018/02/04

（3）ページ内のテキストを用いて予め大別した 6 つのカテゴリに分類

Onion ドメインを大まかに 6 つのカテゴリに分類する．Web サイトのカテゴリ分類する際よく使用される手法としてナイーブベイズが挙げられる．このアルゴリズムは Web サイトのカテゴリ分類のほか分書分類やスパムメールのフィルタなどに用いられる．

文書 *doc* が与えられたときカテゴリ *cat* である事後確率は $P(cat|doc)$ は式 1 以下のように表せる．

$$P(cat|doc) = \frac{P(cat)P(doc|cat)}{P(doc)} \propto P(cat)P(doc|cat) \quad (1)$$

分書 *doc* は *bag-of-words* で単語の集合体として表すことができ，単語間が独立であると仮定すると式 2 以下のように計算することができる．

$$\begin{aligned} P(cat|doc) &= P(word1 \wedge \ldots \wedge word_k|cat) \\ &= \prod_i P(word_k|cat) \end{aligned} \quad (2)$$

また$P(cat|doc)$ は掛け算部分がアンダーフローを起こす可能性があるため対数をとって掛け算を足し算化する．しかし，このままだと訓練データに一つもない単語が表れた場合，$P(cat|doc)$ が 0 となってしまい計算ができなくなってしまう．これをゼロ頻度問題と呼ばれている．そこでよく使われる解決手法として，単語の出現回数に 1 を加えるラプラススムージング（Laplace Smoothing）が挙げられる．よって一般的なナイーブベイズは式 3 以下のように示される．

$$\begin{aligned} P(word_i|doc) &= \frac{T(cat, word_i) + 1}{\sum_{word' \in V}(T(cat, word') + 1)} \\ &= \frac{T(cat, word_i) + 1}{(\sum_{word' \in V} T(cat, word')) + |V|} \end{aligned} \quad (3)$$

しかし，この場合のテストデータは必ず教師データのカテゴリに属する結果になってしまうが，Dark Web は多種多様なカテゴリが存在するため上記の式は相応しくない．よって，ゼロ頻度問題が発生したようなときは，分類不可とす

Vol.2018-IOT-43 No.10
Vol.2018-SPT-30 No.10
2018/9/28

るためラプラススムージングを適用しないで分類する．

教師データの例を表 2 に示す．単語を定義する際，例えば“btc“など複数のカテゴリにも属してしまうような単語は除くようにし，教師データ数は 200 以上とした．

表 2　カテゴリに属するキーワード例

hacking	drug	develop	porn	news	casino
rats	drug	debian	sex	terror	casino
backdoor	marihuana	apt	sexual	syria	roulette
malware	cannabis	ubuntu	teen	china	slots
ddos	narcotic	program	girl	legal	poker
booter	meth	package	porn	news	blackjack
hacker	herb	oss	adult	blog	baccarat
zero-day	mema	develop	erotic	rss	trump
threat	gram	centos	hentai	chat	craps
spy	pill	source. list	asian	form	game

ナイーブベイズ分類結果を図 3 に示す．全体の約半分が分類不可である None という結果である．一番少ないのが OS でありこれは，パッケージ情報やリファレンスである．対して一番多いのが news でありこれは，ニュース記事や政治的主張をするサイトである．

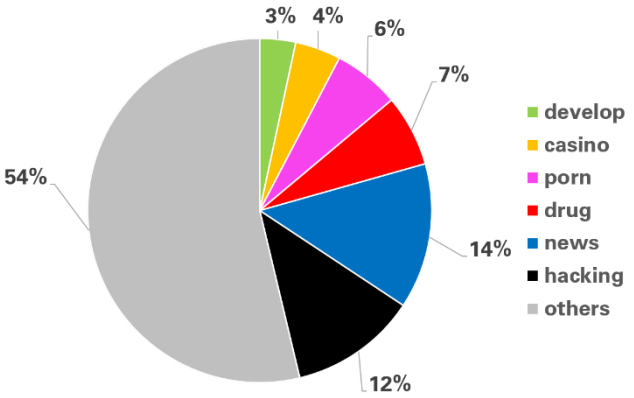

図 3　ナイーブベイズ分類結果

（4）ページの内のハイパーリンクから有向グラフを作成する

まず，ダウンロードしたページに対して“.onion”が存在するようなハイパーリンクの解析を行った．解析結果を図 4 に示す．5,375 のアドレスのうち自身を除いて 1 件以上リンクがあるアドレスは 1,086 しかなく約 80%以上のアドレスは 1 件もリンクを持たないことが判明した．また，約 99%のアドレスは 100 件以下のリンク数である．

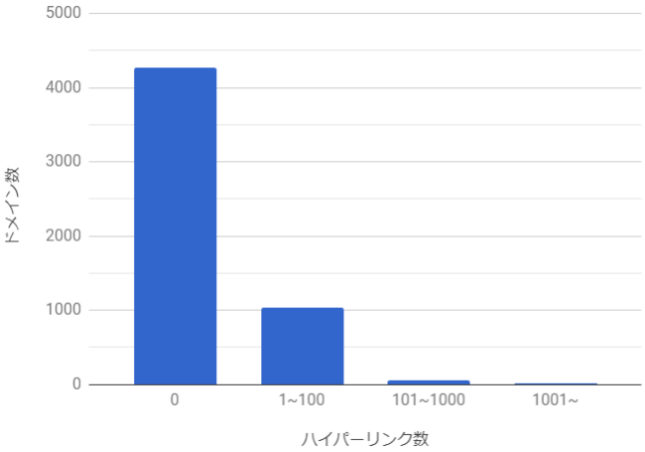

図 4　ハイパーリンク解析結果

ノードを Onion ドメイン，リンクされているドメインに対して方向性を持つようなグラフを作成する．結果を図 5 に示す．

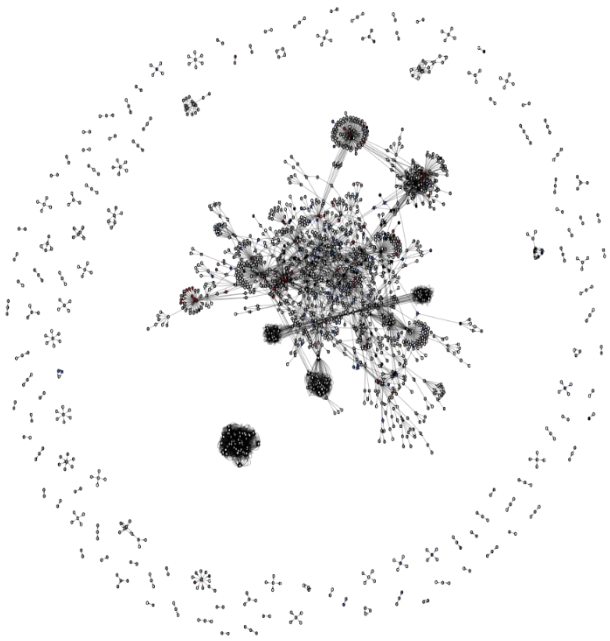

図 5　有向グラフ

5.2　（5）作成した有向グラフにカテゴリ分類結果を反映

5.3 で述べた 6 つのカテゴリに分けた結果をグラフに反映する．その際，視覚的に分かりやすいよう各カテゴリに対して色を定義し，ノードに色を付与する形式でグラフを作成する．作成したグラフサイズが大きいため一部拡大し図 6 に示す．

このようなグラフを作成することで，Dark Web の一端をネットワークグラフの様に表現し，本研究の達成目標の一つである Dark Web のカテゴリごとのつながりを可視化することができた．また新たな Onion アドレスの発見とコンテンツの推測等に利用できると考えている．

Vol.2018-IOT-43 No.10
Vol.2018-SPT-30 No.10
2018/9/28

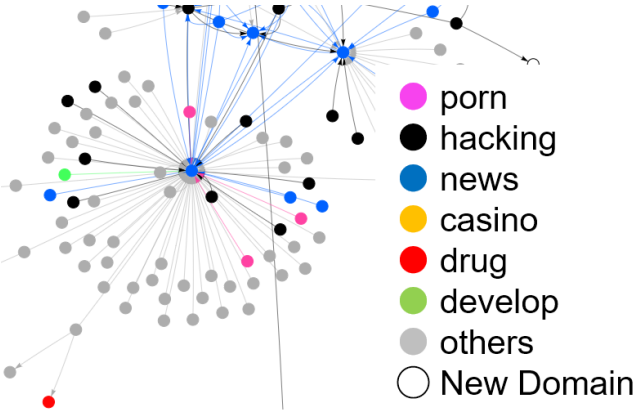

図 6　カテゴリ有向グラフ　拡大

6.　考察

6.1　分類結果について

全体の半数以上が分類不可となってしまったのは改善の余地が大きくあると思われる.原因として 2 つ考えられる. 1 つ目は，キーワードとして定義した単語は全て英語で指定した点にあると考えられる. Dark Web は多くの言語は英語だが，英語以外にもロシア語・中国語・フランス語・日本語など多数の言語で書かれている．よって英語以外の Web サイトは分類できなかったと思われる．これについては全ての言語においてキーワードを定義するのは現実的に実施するのは困難である．2 つ目は，カテゴリに属するキーワードに不足があり正しく分類できなかった点である．カテゴリを定義する際，ランダムに 300 サイトを手動で閲覧し特徴的なキーワードを抽出し，さらにカテゴリに含まれそうな単語を考え列挙した．一つのカテゴリに対して最低 10 個以上のキーワードを定義した. しかし, 例えば“drug”カテゴリでは薬物をスラングのような言い回しが存在するため全てを列挙することは困難である.

1 つ目の問題点の解決策として翻訳サイトを使い全て英語に翻訳することである．翻訳サイトは言語を自動的に判定するため元のテキスト文の言語に考慮する必要がないのである．2 つ目の問題点の解決策としてできるだけ多くのキーワードを定義する，ということを現状考えている.

分類できた Web サイトを実際に手動で閲覧した結果，高い確率でカテゴリが一致していた．したがって分類できた多くの Web サイトは概ね正しく分類できたと言える.

6.2　コンテンツのつながりについて

今回作成したグラフにはいくつかの特徴がある．ノードの次数に着目した場合とカテゴリに着目した場合の 2 つを述べる.

エッジの接続数が 10 未満の小さな集団がグラフの外周に 400 以上存在する．グラフの中心付近にもいくつかの小さな集団は存在するが，集団内の 1 つ以上のノードは必ずどれかの集団と 1 つ以上の接続がある．よって外周に円状に存在する集団は他の Onion ドメインとのつながりが希薄であり，アクセス数が少なくサイト自体の規模が小さいと考えられる．次にグラフ中心付近の小さい集団のうち集団同士を接続するノードはコミュニティサイトや情報を発信するようなブログである可能性が高い．なぜなら Tor ネットワークには Google や Yahoo のような有名な検索エンジンがいくつか存在していたが, 現在は多くが閉鎖されている. よってインターネット黎明期のようにリンク集や情報サイトによって Web サイトのアドレス知るパターンの方が多いのではないか，と考えられるからだ．最後に 1,000 以上の次数を持つノードの存在である．そのようなノードは 16 個存在し明らかに以上である．またただ 1 つだけ次数が 21880 と異常なノードが存在する. 異常な次数を持つ 16 個の Onion アドレスに対して手動でアクセスした結果，検索エンジンとチャットサイトであることが判明した.

同じカテゴリ同士は固まった集団を形成することが分かった．例として drug の集団を図 7 に示す．別のカテゴリを示したノードもあるが多くが drug のノードであり相互に密に結びついているのが分かる.

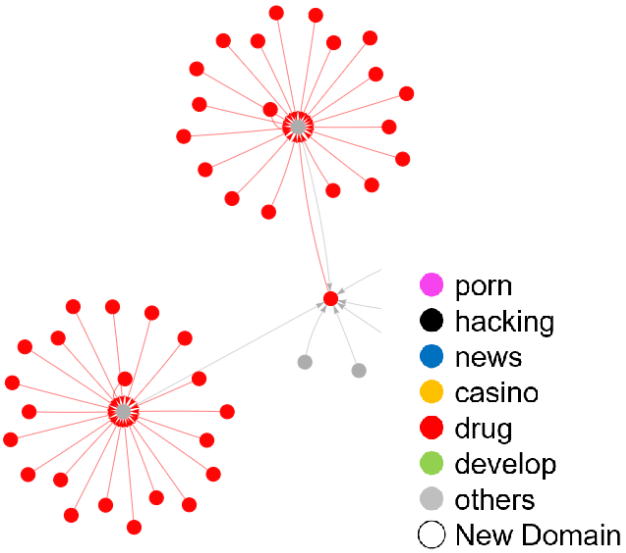

図 7　類似したカテゴリの集団

7.　まとめ

本稿では，Tor ネットワーク上の Dark Web をターゲッ

[13] 小野諒人，神薗雅紀，笠間貴弘，上原哲太郎，“HSDir の Snooping と秘匿サービスへのスキャンを組み合わせたダークウェブ分析システム”，電子情報通信学会，2018 トとしたコンテンツの分類とつながりについて

の分析方法を提案し実装した．Ichidan を用いて 8,291 件の Tor アドレスを収集し，トップページのテキスト文からナイーブベイズのアルゴリズムを用いてカテゴリ分類した．分類したカテゴリは違法とされる Web サイトだけでなく，Dark Web に存在するとされる 6 つのカテゴリに分類した．そしてハイパーリンクからカテゴリごとのつながりを可視化した．可視化する際，分類したカテゴリの他に分類不可・新規ドメインも含めノードに色を付与することでカテゴリ同士がどのように接続するかが容易に理解できる．作成したネットワークグラフから Dark Web のコンテンツの同士のつながりと特徴を明らかにした．

参考文献

[1] BERGMAN Michael K, “The Deep Web: Surfacing Hidden Value”, Journal of Electronic Publishing, 2001
[2] 日経TECH, “ダークウェブ”, http://tech.nikkeibp.co.jp/atcl/nxt/column/18/00178/040600011/
[3] ASCII, “警察情報を漏えいさせた匿名通信システム「Tor」の秘密”, http://ascii.jp/elem/000/000/588/588241/
[4] 東京新聞, “闇サイト事件１０年　犯罪情報　ネット深く”, http://www.tokyo-np.co.jp/article/national/list/201708/CK2017082402000235.html
[5] Tor Metrics, “Servers - Tor Metrics”, https://metrics.torproject.org/networksize.html
[6] Tor Blog, “Ethical Tor Research: Guidelines”, https://blog.torproject.org/ethical-tor-research-guidelines
[7] THE ZRRO / ONE, “ダークウェブをスパイする「Tor HSDir」が100以上見つかる”, https://the01.jp/p0002855/
[8] THE ZERO / ONE, “サンパウロの大学がダークウェブを覗き見？ Torリレー運営から追放される”, https://the01.jp/p0005695/
[9] Ichidan, http://ichidanv34wrx7m7.onion/
[10] Shodan, https://shodan.io/
[11] 情報セキュリティ, “増加するインターネット接続機器の 不適切な情報公開とその対策”, https://www.ipa.go.jp/files/000052712.pdf
[12] Forbes Japan, ダークウェブに14億件の個人データ流出、有名ポルノサイトも, https://forbesjapan.com/articles/detail/18912
[13] Ahmed T. Zulkarnine, Richard Frank, and Bryan Monk, “Surfacing collaborated networks in dark web to find illicit and criminal content”, Intelligence and Security Informatics, 2016
[14] 小野諒人, 神薗雅紀, 笠間貴弘, 上原哲太郎, “HSDirのSnoopingと秘匿サービスへのスキャンを組み合わせたダークウェブ分析システム”, 電子情報通信学会, 2018