IDN ホモグラフ攻撃の大規模実態調査: 傾向と対策

鈴木 宏彰[1]　森 達哉[1]　米谷 嘉朗[2]

概要：国際化ドメイン名 (IDN) は，ドメイン名として漢字，ひらがな，ハングル等の非 ASCII 文字を，Unicode で表現する仕組みである．IDN で利用可能な文字集合の中には，異なる文字符号が割り当てられているにもかかわらず，見た目が非常に似た文字が存在する．例えばラテン文字の ‘a’(U+0061) とキリル文字の ‘ а ’ (U+0430) はその一例である．このような異なる文字の外形的類似性を利用し，URL 偽装によるフィッシングを行う **IDN ホモグラフ攻撃**が知られている．本研究は代表的な TLD として，`.com`，`.net`，および`.jp` のドメイン名を対象とし，IDN ホモグラフ攻撃に使われている，あるいはオリジナルのドメイン名の所有者がホモグラフ攻撃対策として防衛的に設置した可能性があるドメイン名の大規模な調査を行う．抽出したドメイン名を持つウェブサイトを分析し，誰が，どのような目的で登録・運用しているかを調査する．また，これまでに報告されてこなかった，CJK 統合漢字で定義される文字集合内に存在する外形的に似た文字（ホモグリフ）を使った IDN ホモグラフ攻撃の可能性を示し，最後にユーザビリティを損なわない IDN ホモグラフ攻撃対策技術について論じる．

キーワード：　国際化ドメイン名，IDN ホモグラフ攻撃，フィッシング

1. はじめに

ドメイン名を表現する文字集合は，大文字と小文字を区別しないアルファベット，数字，ハイフンなどの ASCII 文字の部分集合からなる．ドメイン名を表現するための文字集合を拡張し，例えば日本語，中国語，ロシア語等，ASCII 文字では表現できない文字集合を利用できるようにするメカニズムは国際化ドメイン名（IDN）として知られている．1996 年に Duerst がインターネットドラフトとして提案した後 [1]，Internationalizing Domain Names in Applications (IDNA) [2] として技術標準がまとめられた．今日利用されている多くのブラウザが IDNA に対応している．

IDN では Unicode で表現される様々な文字をドメイン名として利用することができる．例えば，`http://`早稲田大学`.jp` や `https://`日本レジストレーションサービス`.jp/` のように，人間にとって覚えやすく，中身を理解しやすい IDN ドメインを用いた URL を，テレビ放送，ラジオ放送，ウェブページ，PDF などの電子媒体，書籍などでユーザに提示することができる．

IDN として利用可能な文字集合には視覚的に酷似した文字が含まれる．このため，既存のドメイン名と外形的特徴が酷似しているものの，実際には異なる符号化文字で構成される IDN ドメイン名の作成が可能である．例えば ASCII 文字で構成される文字列 `facebook` に対して，Unicode 文字である ’`LATIN SMALL LETTER E WITH ACUTE`’ (U+00E9) すなわち ‘é’ を使い，`facébook` を構成することができる．元の文字列と，それに似た文字で構成した文字列の外形的特徴はきわめて似ていることがわかる．‘é’ は ‘e’ にアキュート・アクセントを付した文字であり，フランス語，イタリア語，オランダ語，スペイン語，ポルトガル語等欧州圏で広く使われる．

上述のような外形的に似た文字を悪用することで，既存のドメイン名に対して視覚的に見分けがつきにくい IDN を登録し，攻撃者が用意した web サイトにユーザーを誘導する攻撃が考えられる．そのような攻撃は **IDN ホモグラフ (homograph) 攻撃**と呼ばれ，フィッシングなどに利用される．IDN ホモグラフ攻撃の問題は古くから指摘されており，2001 年には Gabrilovich と Gontmakher が IDN ホモグラフ攻撃の脅威を例示している [3]．IDN ホモグラフ攻撃は現実的には大きな脅威として考えられてこなかったが，IDN の登録・運用実績が増え，ブラウザやアプリでの対応が進んだことから，IDN ホモグラフ攻撃による脅威が現実的なものになってきた [4, 5]．

本研究では IDN ホモグラフ攻撃に利用されている可能性があるドメイン名すなわち**潜在的ホモグラフドメイン名**

[1] 早稲田大学 基幹理工学研究科
[2] 株式会社日本レジストリサービス

の実態を大規模に調査する．我々の狙いは，(1) 潜在的ホモグラフドメイン名を**誰が，どのような目的**で登録・運用しているのかを明らかにすること，および (2)IDN の利点を損なうことなく，IDN ホモグラフ攻撃に対する適切な**対策方法**を考案することにある．

本研究は代表的なトップレベルドメイン (TLD) として，.com，.net，および.jp を分析の対象とする．.com は 1 億以上のレコードを持つ世界最大の TLD であり，.jp は一つの国全体のドメインを司る ccTLD である．これら TLD の DNS ゾーンファイルに登録された全ドメイン名を分析対象とする．分析対象としたそれぞれの TLD に登録されたドメイン数は 136,369,877 個 (com), 14,372,833 個 (net), 1,540,767 個 (jp) であった．はじめにそれぞれの TLD に登録されているすべてのドメイン名から IDN を抽出する．次に Alexa Top Sites [6] において，各 TLD における上位 1,000 のドメイン名を抽出する．これらのドメイン名集合と前記の IDN の集合と Alexa ドメイン名集合からそれぞれ 1 つずつドメイン名を抽出し，文字列として比較する．文字列比較の差分が 1 文字のみとなり，かつ差分となった 2 つの文字の外形的特徴が似てた場合，対応する IDN を潜在的ホモグラフドメイン名として抽出する．文字列比較において，参照候補のフィルタリングと編集距離による評価を適用することにより，高速かつ高精度に潜在的ホモグラフドメイン名の抽出が可能である．

次に抽出した潜在的ホモグラフドメイン名を良性，悪性，要注意の 3 カテゴリーに分類する．それぞれ良性は，元の Alexa 掲載のドメイン名の所有者が防衛的な目的で取得したと考えられる IDN の集合，悪性は元の Alexa 掲載のドメインとは異なる所有者が取得した潜在的ホモグラフ IDN ドメイン名のうち，悪性判定がされた IDN の集合，要注意は元の Alexa 掲載のドメインとは異なる所有者が取得した潜在的ホモグラフ IDN ドメイン名のうち，特に悪性な挙動は示さない IDN の集合である．さらに悪性と判定されたドメイン集合に対して，そのドメインを検索してヒットしたウェブサイトを分析し，ウェブサイトで提供されているサービスを調査した．

既存文献，および上記で分析対象とした IDN ホモグラフは元が ASCII 文字ドメインに対して，IDN でホモグラフを構成するモデルであったが，原理的には元々 IDN であるドメインに対してもホモグラフを構成することが可能である．例えば，CJK 統合文字で構成される IDN に対し，同じく CJK 統合文字で構成される IDN ホモグラフを作成することができる．本研究ではそのような例のひとつとして，日本語 IDN に対する潜在的ホモグラフドメイン名とその構成方法を調査する．例えば，○○工業.jp に対して○○エ業.jp はホモグラフドメイン名である（漢字の「工」とカタカナの「エ」の外形的類似性を利用）．

本研究の貢献は以下の通りである．

- IDN を効率的に検出する手法を提案した．画像処理を用いて IDN ドメインを検出する既存手法 [4] と比較して，211 倍の高速化を達成した．
- Alexa Top 1,000 を対象とした，潜在的ホモグラフドメインを分析した結果，約 1,600 のホモグラフドメイン名を発見し，その大部分が悪性，もしくは要注意な IDN ホモグラフであることを明らかにした．
- ひらがな，カタカナ，および CJK 統合漢字の文字を対象とし，既存 IDN に対するホモグラフ攻撃の実現可能性を評価した．
- IDN が持つユーザビリティを損なわず，かつユーザに対して潜在的ホモグラフドメインに対して効果的な注意をうながす対策技術を提案した．

本章の構成は以下の通りである． 2 章では，本研究の背景知識として IDN の概要を示す． 3 章では，本研究で収集したデータ，およびデータの分析方法を示す． 4 章にデータを分析して得られた結果を示す． 5 章では本研究の制約事項，ならびに IDN ホモグラフ攻撃への対策を議論し，具体的な対策の Proof of Concept を示す． 6 章で関連研究を示し， 7 章はまとめである．

2. IDN の概要

本章では，国際化ドメイン名 (IDN) の概要を説明する．DNS では ASCII 文字における英数字とハイフンからなる Letter-Digit-Hyphen (LDH) と呼ばれる文字集合を使うことができる．IDN はドメイン名として，LDH 以外の文字集合を利用できるようにする仕組みであり，IDN の導入により，Unicode で符号化された文字で表現可能なきわめて多種多様なドメイン名を登録・公開することができる．IANA は TLD レジストリ事業者に向けて，IDN の登録・運用に関するガイドラインを提供している [7]．ガイドラインでは，IDN に用いることができる文字は，TLD レジストリ事業者が明示的に許可した文字集合に限定することを推奨している．IDN として利用可能な文字集合はトップレベルドメイン (TLD) ごとに異なっており，IDN tables として管理されている [8]．

ドメイン毎に IDN として利用可能な文字集合に制限を設けることにより，IDN ホモグラフ攻撃の脅威を緩和することができる．例えば JP ドメインの場合，IDN として利用な文字は LDH，ひらがな，カタカナ，漢字に限定されている．したがって，JP ドメインの配下のセカンドレベルドメインでは LDH に含まれるアルファベットに似た文字（キリル文字やギリシャ文字）を使った IDN ホモグラフの登録はできない．

DNS の仕様・実装で利用可能な文字空間は LDH に限られるため，Unicode で表現される IDN を LDH に変換する仕組みが必要である．Punycode は Unicode で書かれた文字列を LDH に変換する文字符号化方式であり，RFC 3492 [9]

で定義されている．Punycode で変換された文字列を IDN で使うときには，変換された文字列の先頭にプレフィックスとして xn-- を付加する．例えば IDN の一例である早稲田.jp における「早稲田」を Punycode で変換すると 0gvz35a4wd となり，IDN としては xn--0gvz35a4wd.jp となる．

IDN の扱いはブラウザの実装ごとに異なる．特にユーザインタフェースとしては，アドレスバーにおける IDN の表示方法が問題となる．2017 年 4 月にウェブ上のブログ記事 [5] によって IDN ホモグラフ攻撃による脅威が広く喧伝された後，多くのブラウザベンダが IDN の表示に関する実装を変更した．具体的には Firefox や Chrome においては，IDN を構成する文字列中に複数のスクリプト（文字集合）に属する文字が混在している場合に，ユニコードではなく Punycode で IDN を表示する変更が行われた [10,11]．特にラテンスクリプト，キリルスクリプト，ギリシャスクリプトが組み合わさった場合は Punycode で表示がされるため，先にあげた facébook のケースでは Punycode である xn--facbook-dya が表示される．複数スクリプトであっても，ラテンスクリプトと CJK イデオグラフ（日本語，中国語，韓国語で使われる文字集合）が混ざっている場合はユニコードの IDN を表示する．

前述のブラウザによる対策は IDN ホモグラフ攻撃の一時的な対策にはなるものの，強制的に Punycode で表示された場合，原因の所在がわかりにくくなる問題がある．すなわち，ユーザはブラウザに入力されたドメイン名がホモグラフ攻撃であることに気が付かないため，再び同じドメイン名のサイトを訪れてしまうリスクがある．また， 4.5 節で示すように，ホモグラフ攻撃が成立するのはラテンスクリプト，キリルスクリプト，ギリシャスクリプトの組み合わせのみならず，CJK イデオグラフに属する文字の組み合わせによっても実現可能である (例えば前述した漢字の「工」とカタカナの「エ」)．この場合，現状のブラウザの対応ではユニコードで IDN が表示されるため，IDN ホモグラフ攻撃の可能性を見逃すリスクがある．

3. データ・分析方法

本章では分析のために収集したデータ，および分析方法の詳細を示す．

3.1 データ

3.1.1 攻撃対象ドメイン

IDN ホモグラフ攻撃は，正規のドメインと視覚的に類似したホモグラフドメイン名にユーザを誘導することで成立する．したがって，IDN ホモグラフ攻撃の標的となり得るドメインは，広く認知されている有名なドメイン名となる可能性が高いと想定する．そのようなドメイン名として，Alexa Top Sites [6] を利用する．本研究では Alexa Top Sites に掲載されたドメインのうち，特に .com，.net を TLD とするドメイン名で，かつセカンドレベルドメイン名の長さが 5 以上のドメイン名を上位から順に 1000 個を抽出した．ドメイン名の長さを 5 以上とした理由は次節で述べる．

表 1　.com，.net，.jp の DNS ゾーンファイルの情報

TLD	ドメイン名数	IDN 数	取得日
.com	136,369,877	1,000,823	2018/06/26
.net	14,372,833	216,997	2018/06/18
.jp	1,540,767	91,180	2018/07/10

3.1.2 DNS ゾーンファイル

現在登録されているドメイン名に潜在的ホモグラフドメイン名が存在するかを調べるために，DNS ゾーンファイルを用いる．DNS ゾーンファイルとは，DNS ゾーンを記載したテキストファイルであり，DNS ゾーンは管理責任が単一の管理組織に委任されている DNS 上の名前空間である．今回は代表的な DNS ゾーンとして .com，.net，.jp の 3 つに着目する．DNS ゾーンファイルには，ドメイン名と IP アドレスの対応 (A レコード) や，ドメイン名とネームサーバ (NS レコード) 等，ゾーン内で定義される DNS レコード情報が記載されている．表 1 に各 DNS ゾーンファイルに掲載されていたドメイン数，IDN 数，およびゾーンファイルの取得日を示す．

Alexa Top Sites に含まれるドメイン名はすべて ASCII 文字で構成されていた．したがって，Alexa Top Sites ドメイン名集合に対する潜在的ホモグラフドメイン名を調査する場合，潜在的ホモグラフドメイン名に使用される文字としては，例えばラテン文字の補助やギリシャ文字で定義されるアキュート・アクセント (アルファベットアクセント符号が付加された文字）などが候補となる．一方，前述したように .jp で利用可能な文字集合には，上記のようなラテン文字のアルファベットに外形的性質が近い文字が含まれていない．従って，本論文では .com，.net の IDN を合計した 1,217,820 個のドメイン名の中から Alext Top Sites に対する潜在的ホモグラフドメイン名を調査し，さらに .com，.net，.jp に含まれていた合計 1,309,000 個の IDN の中から，日本語 IDN に対する潜在的ホモグラフドメイン名を調査する．

3.1.3 WHOIS データベース

検出した潜在的ホモグラフドメイン名の登録者情報を得るために，WHOIS データベースを利用した．WHOIS とは，ドメイン名に関連する情報を収集するためのプロトコルである．WHOIS データベースに登録される情報は，ドメイン名を登録したレジストラ，ネームサーバ，失効期限などである．これらの情報は，ドメインの取得状況の確認や，ドメイン名が不正利用された場合のコンタクト情報の入手などの目的で利用される．WHOIS データベースに登録された情報の収集には標準的な WHOIS コマンドを利用

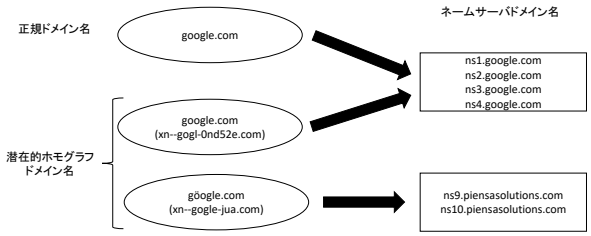

図 1　google.com に対する潜在的ホモグラフと NS レコード

し，クエリに対する応答をテキスト解析した．

抽出した潜在的ホモグラフドメイン名が，あるドメインに対する IDN ホモグラフ攻撃に対する防衛の目的で取得された場合，WHOIS に登録される情報としては，元のドメインと同じレジストラによる登録がなされ，さらに同じ NS レコードを持つ可能性がある．外部の企業に防衛目的のドメイン名の取得を委託するようなケースもある．その場合，特定したレジストラ，NS レコードの統計を調査することにより，そのようなドメイン登録業者を特定することが可能である．

図 1 に google.com に対する潜在的ホモグラフドメイン名を抽出し，NS レコードによって分類した例を示す．

3.1.4　VirusTotal

潜在的ホモグラフドメイン名の中で，悪意のある挙動をするウェブサイトに利用されているかどうかを調査するために，VirusTotal (VT) を利用する．VT はオンラインスキャンサービスの一つであり，URL を入力すると複数のアンチウィルスソフトを使用して，そのウェブサイトをスキャンした結果を出力するサービスである．VT を用い，抽出した潜在的ホモグラフドメイン名が悪性であるかどうかを判定できる．また，悪性判定の結果を分類することができる．今回は「マルウェア配布サイト」,「フィッシングサイト」,「悪性サイト」の 3 分類を採用する．悪性サイトにはエクスプロイトや悪性なコードなどが含まれている．

3.2　潜在的ホモグラフドメイン名の検出方法

3.2.1　編集距離

潜在的ホモグラフドメイン名を検出するアプローチとして，二つの文字列間の類似度を示すメトリクスの一つである編集距離 (Levenshtein 距離) を用いる．編集距離は，2 つの文字列がある場合に，片方の文字列に対して置換，挿入，削除からなる 3 つの編集処理を繰り返して，他方の文字列と一致するために必要な，最小の操作回数によって定義される．編集距離は動的計画法によって算出することができる．

3.2.2　潜在的ホモグラフドメイン名の検出手順

以下に編集距離を用いて，潜在的ホモグラフドメイン名を検出する手順を示す．手順は以下の 3 つのステップからなり，それぞれのステップを各 DNS ゾーンファイルに適用する．

ステップ 1: Alexa Top Sites のドメイン名集合より，特にセカンドレベルドメインの文字列長が 5 以上のドメイン名を抽出する．ドメイン名の長さが短い場合，編集距離が 1 や 2 に該当するパターンが指数的に増えるため，このような制限を設けた．次に上記のドメイン名集合に含まれるすべてのドメイン名と，各 DNS ゾーンファイルに含まれるすべての IDN のペアについて，編集距離を計算する．なお，編集距離を計算する前に IDN を Punycode 表記から Unicode 文字列に変換する．これにより，Unicode 文字列同士で編集距離を算出することができる．また，編集距離を計算するペアは文字列長が同じものに限定することで，組み合わせを大幅に減らすことができる．編集距離の算出結果が 1 もしくは 2 であったとき，対応する IDN を潜在的ホモグラフドメイン名として抽出する．

ステップ 2: ステップ 1 で抽出したペアにおいて差分となった置換文字に着目する．置換文字を手動で検査し，形状が似ているものを採用，似ていないものを除外した．ここで置換文字として近いと判定したものを類似文字リストとして登録することにより，2 回目以降の分析ではそのようなリストを参照することで処理の自動化が可能である．

ステップ 2 の処理は画像処理によって自動化することも可能であるが，**計算的に算出する画像の類似性と人間が認知する類似性は必ずしも一致しない**ため，今回は人間の目視によるチェックを行った．後述するように，今回対象とした約 120 万の IDN に対してステップ 2 を実行した際に要した時間は約 5 時間であった．視認によって類似性をチェックするプロセスを複数ユーザで実施すること，および機械学習などの手法で補強することは今後の課題である．

ステップ 3: ステップ 2 の後，編集距離が 1 または 2 となり，かつ差分となった置換文字が類似文字リストに入っていない場合，該当する IDN を正規のドメインに対する潜在的ホモグラフドメイン名として検出する．

3.3　潜在的ホモグラフドメイン名の分類

潜在的ホモグラフドメイン名をホスト名とした URL に対して HTTP GET メソッドでアクセスし，リダイレクトなどを経て最終的に到達したサイトのドメイン名を調査した．到達先のドメイン名や HTTP ステータスコードにより，潜在的ホモグラフドメインを以下のように分類することができる．

良性ホモグラフ

ドメイン名はアクセスした IDN のままであるが (Unicode 表記あるいは Puny 表記のどちらでも)，図 1 に示した g

o ogl e.com (xn--gogl-0nd52e.com) の例のように[*1]，ネームサーバがオリジナルのドメインである場合，あるいはアクセスした先のドメイン名がオリジナルのドメイン名にリダイレクトされる場合は防衛目的で取得されたドメイン名と考えることができる．そのようなドメイン名を「良性ホモグラフ」と分類する．

悪性ホモグラフ

潜在的ホモグラフドメイン名に HTTP アクセスしてリダイレクトされた結果，オリジナルのドメインとはまったく異なるドメイン名のサイトに到達した場合，VirusTotal による検査を行う．検査の結果，悪性判定された場合は IDN ホモグラフ攻撃に使用されている可能性が高い．そのようなドメイン名を「悪性ホモグラフ」と分類する．

要注意ホモグラフ

オリジナルのドメイン名とは異なるサイトにリダイレクトされるが，VirusTotal の検査結果は特に悪性ではない場合，転売目的で取得されたドメイン名である可能性が高い．あるいはアクセスした際の HTTP ステータスコードでエラーが返された場合，防衛目的で取得した所有者が意図的にアクセスできないようにしているか，転売目的で取得したがアクティベートしていない可能性がある．いずれの場合も現時点での脅威は低いと考えられる．これらのドメイン名は，手動で判断する必要がある，要注意なホモグラフである．そのようなドメイン名を「要注意ホモグラフ」と分類する．

3.4 悪性ホモグラフドメイン名の分布調査

悪性ホモグラフドメイン名に対して Google 検索を行い，該当するドメイン名を含むウェブサイトの分布を調査する．調査には Selenium と Chrome ヘッドレスブラウザを用いる．検索にヒットしたサイトの説明文の中に，該当ドメイン名を含むサイトの URL を保存する．また，1 回のクエリに対するトップ 60 件の応答を調査対象とする．

3.5 IDN 同士のホモグラフドメイン名

既存の IDN に対して，ホモグラフドメイン名を登録することが可能である．例えば 早稲田大学.jp に対して，早稲田大学.jp (xn--pss25cv9p80po3f.jp) を登録することができる．ここで「稲」(U+7A3B) は「稲」(U+7A32) の旧字体であり，いずれも CJK 統合漢字に含まれるため，IDN として利用可能である．

IDN において利用可能なひらがな，カタカナ，および CJK 統合漢字において，外形的に似ている文字のペアを抽出する．手始めに Unicode Consortium が提供している外形が近い文字のペアのリスト confusable.txt [12] に含まれている文字集合と.com，.net，.jp においてそれぞれの TLD に対応する IDN table [8] で定義されている利用可能な日本語の文字集合のマッチングを行う．ただし，後に議論するように confusable.txt でカバーされている外形が類似している CJK 統合漢字は全体のごく一部であると考えられる．そこで本研究では日本の常用漢字表に旧字体が収録されている 364 組の新字体，旧字体のペア [13] を対象とし，IDN table で利用可能な CJK 統合漢字を抽出する．旧字体の一部は CJK 互換漢字 (CJK Compatibility Ideographs) に収録されているため，IDN として使うことができないが，それ以外では利用可能な新字体・旧字体のペアが存在し，外形が似ているものがある (例えば先の「稲」と「稲」).

[*1] フォントによって IDN の見栄えが異なることに注意.

表 2 検出した潜在的ホモグラフドメイン名の統計

TLD	編集距離	ドメイン数
.com	1	1,246
	2	302
.net	1	85
	2	5
合計	–	1,638

4. 分析結果

本章では 3 章で説明したデータと分析方法から得られた結果を示す．

4.1 検出した潜在的ホモグラフドメイン名

表 2 に検出した潜在的ホモグラフドメイン名の統計を示す．各 TLD において編集距離が 2 のドメイン名よりも 1 のドメイン名のものが多くなっていることが分かる．これはホモグラフ攻撃ができる限り正規ドメインと外見を似せることを目的としていることを考えると妥当な結果であると考えられる．

表 2 に示した潜在的ホモグラフドメインを計算する際のステップ 1 に要した合計時間は約 30 分 (1,767 秒) となった．また，ステップ 2，3 に要した時間は約 5 時間であった．Liu らによる既存研究 [4] では画像処理によってホモグラフドメイン名を検出する方法を提案している．ホモグラフの検出にはメモリ 4GB を搭載した計算機を利用し，合計で 102 時間を要したことを報告している．また，Liu らの報告したホモグラフ IDN 検出数は 1,516 であり，同程度のホモグラフ IDN を検出できていることがわかる．したがって，我々のアプローチは Liu らの方法と比較して，$102/(0.5+5) = 18.5$ 倍の高速化を達成している．さらにステップ 2，3 は一度だけ実施すれば良いため，その時間コストを除くと約 211 倍の高速化が達成できたといえる．

なお細かな条件の違いとして，Liu らの分析した IDN の数がおよそ 140 万個であったのに対し，我々が分析対象とした IDN 数は 120 万個であること，また，Liu らは Alexa Top Sites のセカンドレベルドメイン名を上位 1,000 個使用したのに対し、本論文では各 TLD で上位 1,000 個のド

表 3　潜在的ホモグラフドメイン名が多く存在するドメイン名

順位	ドメイン名	個数	Alexa ランキング
1	google.com	115	1
2	facebook.com	91	3
3	amazon.com	87	11
4	apple.com	60	73
5	instagram.com	53	12
6	binance.com	42	844
7	youtube.com	40	2
8	coinbase.com	39	918
9	icloud.com	37	358
10	paypal.com	31	64

表 4　潜在的ホモグラフドメイン名の分類結果

TLD	良性	悪性	要注意
.com	79	256	1,213
.net	0	7	83
合計	79 (4.8%)	263(16.1%)	1296 (79.1%)

メイン名，合計 2,000 個のドメイン名を利用している違いが挙げられる．

4.2　ホモグラフが多数生成されたドメイン名

表 3 に潜在的ホモグラフドメイン名が多数生成されたドメイン名の上位 10 を示す．多くの場合は Alexa ランキングにおいても上位のドメイン名だったが，`binance.com` および `coinbase.com` はいずれも仮想通貨の取引所のサイトであり，Alexa ランキングは必ずしも上位ではなかった．仮想通貨取引所のサイトはいずれもフィッシングや Exploit などの攻撃を受けやすく，攻撃者にとっても，防御側にとっても IDN ホモグラフ攻撃の対象として注目されていることがみてとれる．取得された IDN ホモグラフの分類は次節以降に示す．

4.3　潜在的ホモグラフドメイン名の分類結果

表 4 に潜在的ホモグラフドメイン名を分類した結果を示す。まず，良性と判定されたホモグラフドメイン名は意外に少ないことがわかる．79 件の内，アクセス先のサイトがオリジナルのドメイン名のサイトと同一であったケースは 71 件であった．また，6 件は HTTP アクセスができず，2 件はエラーが発生した．

悪性判定されたホモグラフドメイン名は全体で 263 個存在した．この内，オリジナルのドメイン名を持つ Web サイトとは異なるサイトに到達したケースが 182 個．そのうち，クライアントエラーなどが起きずアクセスに成功したケースが 161 個であった．これらの 161 個の悪性ホモグラフドメイン名が現時点で特にリスクが高いものである．検出された悪性ホモグラフドメイン名の内訳は，マルウェア配布サイトが 9 個，フィッシングサイトが 170 個，悪性サイトが 84 個という結果であった．フィッシング攻撃として利用されるケースが多いだけではなく，マルウェア配布や悪性サイトに使われることがみてとれる．**表 5** にフィッシングサイトの標的となった上位 10 のドメイン名

表 5　フィッシングサイトと判定されたホモグラフドメイン名が多いドメイン名

順位	ドメイン名	フィッシングサイト数	割合 (%)
1	facebook.com	19	20.1
2	google.com	17	14.8
3	amazon.com	13	14.9
4	microsoft.com	11	36.7
5	coinbase.com	10	25.6
6	netflix.com	10	52.6
7	icloud.com	10	27.0
8	apple.com	10	16.7
9	binance.com	9	21.4
10	instagram.com	8	15.1

図 2　`facebook.com` に対してフィッシングサイトと判定されたホモグラフドメイン名の例

表 6　要注意と判定されたドメイン名のネームサーバドメイン名

順位	ネームサーバドメイン名	出現回数	概要
1	domaincontrol.com	325	レジストラ・パーキングサービス有り
2	googledomains.com	170	Google が提供するドメイン登録サービス
3	registrar-servers.com	135	レジストラ
4	cloudflare.com	67	CDN
5	name.com	35	レジストラ
6	above.com	29	レジストラ，ドメインの売買可能，パーキングサービスあり
7	digitalocean.com	28	クラウド事業者
8	markmonitor.com	26	ドメイン管理企業
9	namecheaphosting.com	22	レジストラ
10	cashparking.com	21	Godaddy が提供するパーキングサービス

を示す．フィッシングサイトと判定された割合が高いドメイン名は表 3 と同様，Alexa ランキングで上位にあるドメイン名である．特に `netflix.com` や `micorsoft.com`，`coinbase.com` 等，ペイメントが関わってくるサイトのドメイン名がフィッシングの対象になりやすい傾向にある．参考として，`facebook.com` に対するフィッシングサイトと判定されたホモグラフドメイン名を**図 2** に例示する．

検出した潜在的ホモグラフドメイン名では，要注意と分類されたものが最も多かった．これらのドメイン名のアクセス先を調べた結果，アクセスできなかったものが 527 件，オリジナルのサイトと異なるページに遷移したものが 769 件存在した．その内，エラーが発生せずにアクセスが成功したものが 653 件であった．要注意な IDN に関して，ネームサーバ (NS) を調べた結果を**表 6** に示す．主としてレジストラが多いことがわかる．すなわち，ドメインパーキングで維持されているケースが想定される．その場合，誰もがドメイン名を購入できるため，悪用されるリスクもある．一方で，`googledomains.com` は一般的なドメイン登録サービスを提供するが，Google が所有するドメイン名の管理もしている．このため，Google が関連するドメインの IDN ホモグラフを防衛的に取得している可能性が

表 7　悪性ホモグラフドメイン名の検索にヒットしたサイト

順位	ヒットしたサイト	ヒット数	ウェブページ概要
1	domain-status.com	357	ドメイン名の登録状況を表示
2	farsightsecurity.com	108	セキュリティー企業の記事
3	urlscan.io	64	URL のスキャン結果
4	atoall.com	40	翻訳サービス
5	facebook.com	38	SNS
6	pastebin.com	35	テキスト共有サービス
7	google.com	34	ドメイン名のブラックリスト
8	domainwat.ch	32	WHOIS 情報の調査結果
9	twitter.com	29	SNS
10	raw.githubusercontent.com	17	フィッシング IDN リスト

表 8　IDN に利用可能な新旧字体ペアの登録数

	.com	.net	.jp
旧字合計	8,707	1,637	288
新字合計	81,616	25,326	44,279

ある．markmonitor.com はドメイン名のブランドを保護する企業であり，このドメイン名の NS レコードを持つ場合，同社のサービスを使って自社のドメイン名を保護している可能性が高い．

4.4　悪性ホモグラフドメインの分布状況

抽出した悪性ホモグラフドメイン名がインターネット上にどのように分布しているかを調査するために，当該ドメイン名を Google 検索した結果を利用した．**表 7** に検索のヒット数が高かった上位 10 のドメイン名を示す．主としてドメインの調査や，セキュリティ検査を目的としたサイトが多いことがわかる．一方で，Facebook や Twitter 等，SNS を経由して情報が流通しているケースがあることも見て取れる．これらのサービスは不特定多数のユーザが訪問する可能性があるため，悪性 IDN がもたらすリスクが高い．

4.5　IDN 同士の潜在的ホモグラフドメイン名

はじめに，confusable.txt [12] に含まれている文字集合と.com，.net，.jp においてそれぞれの TLD に対応する IDN table [8] で定義されている利用可能な日本語の文字集合のマッチングを行う．この結果，12 組，合計 25 個の文字が外形的に類似しており，IDN ホモグラフ攻撃に利用できることがわかった．組み合わせとしてはカタカナと漢字のケースが多く，例えば漢字の「卜」(U+535C) とカタカナの「ト」(U+30C8) などがあった．これらの文字を利用した IDN の数はカタカナの「ト」が.com で 15,547 件，.net で 5,756 件，.jp で 5568 件あったのに対し，漢字の「卜」は.com で 27 件，.net で 7 件，.jp で 0 件であった．

次に常用漢字表に旧字体が収録されている 364 組の新字体，旧字体のペア [13] を対象とし，IDN tables で利用可能なペアを抽出した．この結果，269 組の新旧字体のペアが IDN として利用可能であった．これらの文字の登録状況を**表 8** に示す．新字旧字ともに一定量の登録があるが，基本的には新字の登録が多い．それらのドメインに対し，旧字を使った IDN ホモグラフ攻撃が実現可能である．

上述した以外にも，IDN として利用が可能である CJK 統合漢字には非常に多数の文字が存在するため，ホモグラフを構成するペアが存在する可能性がある．JP ドメインに対する IDN tables の元となる「汎用 JP ドメイン名登録等に関する技術細則」[14] では明らかに外形が近い文字は除外されているが，（例えば「〆」U+3006 と U+4E44 で表現されるカタカナの「メ」に近い漢字など）．com に対する中国語ドメインで許可された文字は，JP ドメインが許可している文字集合と比較して，非常の多数の文字の利用が許可されており，かつ明らかに外形が似ている文字の利用が許可されている（例えば U+5C13 と U+5C14 など）．このため，ホモグラフの構成が容易であると考えられる．今回の研究で探索しきれなかったホモグラフを構成する文字の調査や，ユーザが認知する類似性の検証は今後の課題である．

5.　議論

本章では本研究の制約事項，ならびに IDN ホモグラフ攻撃への対策について論じる．

5.1　制限事項

本研究では規模が大きい代表的な TLD として.com，.net，.jp の 3 つに着目し，IDN ホモグラフドメイン名の分析を行った．他の TLD における IDN の実態把握は今後の課題である．また，IDN 間のホモグラフに関しては日本語を対象としたが，他の言語を対象とした IDN の分析も今後の課題としたい．また，今回の分析はある日時におけるスナップショットで行っているが，時間的な変化，とりわけ所有者の変更等に注目した分析は将来の課題である．

潜在的ホモグラフドメイン名の検出においては，置換された文字の類似性チェックを視認によって確認した．対象となる文字を十分に絞り込んだため，時間的コストは大きなものではなかったが，実際の作業は 1 人で行ったため，複数の評価者による検証が望ましい．今後の課題とする．また，画像としての類似性，機械学習を適用することにより，類似文字検出の自動化が可能になると考えられる．Unicode は依然として発展途上の文字符号化方式であり，今後も新たな文字符号が追加されていく．それらの新しい文字に対して類似文字検出が自動化できれば，標準化の過程においても有用な知見を与えることが期待できる．

今回対象とした IDN は英語，日本語等の自然言語の表記体系で使われる文字を対象としたが，Unicode で符号化されている文字としては，絵文字 (emoji) あるいは「その他のシンボル」として分類される特殊な文字がある．これらの文字を IDN として利用できる TLD があり，そうしたケースでは，新たなホモグラフ攻撃が成立する可能性が考えられる．こうした特殊な文字を使った IDN とそのセキュリティ課題に関する調査は今後の課題である．

図 3 文字差分を利用した IDN ホモグラフ攻撃対策のイメージ

5.2 IDN ホモグラフ攻撃への対策

2 章で示したように，現状のブラウザによる IDN ホモグラフ攻撃対策は，ユーザに対する透明性が失われる点，および 3 章で示した IDN 間ホモグラフ攻撃に対して無効である欠点がある．IDN ホモグラフ攻撃の可能性があることを，ユーザビリティを損なうことなく，明示的にユーザに伝えるには，一律に Punycode に強制変換するのではなく，あるドメイン名が潜在的ホモグラフドメイン名であることを検知した上で，元のドメインと潜在的ホモグラフドメインの**差分**を強調するユーザインタフェースの採用が望ましいと考えられる．**図 3** にブラウザ拡張機能として提案手法を実装したイメージを示す．Unicode 表記を残すことでユーザビリティを高めつつ，差分を明示することでユーザがホモグラフドメインであることを意識すれば，そのドメインに対する警戒を高める効果を期待できる．方式の実装，評価は今後の課題であり，オリジナルのドメイン名をどこを起点にするか（Alexa データを使うのか），どのようなアルゴリズムを使うのか，スケーラビリティにどう対応するか等技術的課題が残されている．

6. 関連研究

Liu ら [4] は Alexa ランキングのセカンドレベルドメイン名の上位 1,000 件と 140 万の IDN を画像処理によって比較することで，ホモグラフドメイン名を調査し，結果として 1516 個のホモグラフドメイン名を発見した．また Alexa ランキングのセカンドレベルドメイン名の上位 1,000 件とホモグラフな文字のデータセットを使用して，類似度を示す指標が高いホモグラフドメイン名を作成し，大多数が登録可能であることを示した．

澤部ら [15] は OCR(光学的文字認識) を使用したホモグラフドメイン名の新しい検出方法を提案した．新しく提案された手法では，手動ではなく OCR により，非 ASCII 文字ごとに似ている ASCII 文字の集合を自動的に作成する．そして非 ASCII 文字の部分を対応する集合の ASCII 文字に入れ替えたドメイン名の中に，正規のドメイン名が含まれている場合に，ホモグラフドメイン名として検出する．またこの方法では，新種の形状が似ている文字も発見できることをメリットとして挙げていた．

7. まとめ

代表的な TLD である `.com`，`.net`，および `.jp` を対象とし，IDN ホモグラフ攻撃に使われている，あるいはオリジナルのドメイン名の所有者がホモグラフ攻撃対策として防衛的に設置した可能性があるドメイン名の大規模な調査を行った．この結果，Alexa Top 1,000 を対象とした約 1,600 のホモグラフドメインを発見し，それらの大部分が悪性もしくは要注意なドメインであることを明らかにした．標的となったドメインを分析した結果，著名サイトに加えて仮想通貨取引所が IDN ホモグラフ攻撃対策になりやすいことがわかった．また，ひらがな，カタカナ，CJK 統合漢字を対象とし，IDN 間ホモグラフの実態調査を行った．この結果，常用漢字表における新旧字体のペアの部分集合を利用した IDN ホモグラフ攻撃が可能であることを示唆した．さらにユーザビリティを損なわない IDN ホモグラフ攻撃対策技術を提案した．対策技術の実装と有効性の評価，ならびに本研究で探索しきれていない IDN 間ホモグラフの網羅的な調査は今後の課題である．

参考文献

[1] Internationalization of Domain Names: `https://tools.ietf.org/html/draft-duerst-dns-i18n-00`.
[2] Klensin, D. J. C.: Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework, RFC 5890 (2010).
[3] Gabrilovich, E. and Gontmakher, A.: The homograph attack, *Commun. ACM*, Vol. 45, No. 2, p. 128 (2002).
[4] Liu, B., Lu, C., Li, Z., Liu, Y., Duan, H., Hao, S. and Zhang, Z.: A Reexamination of Internationalized Domain Names: The Good, the Bad and the Ugly, *Proc. IEEE/IFIP DSN 2018*, pp. 654–665 (2018).
[5] Zheng, X.: Phishing with Unicode Domains, `https://www.xudongz.com/blog/2017/idn-phishing/` (2017).
[6] AWS: Alexa Top Sites, `https://aws.amazon.com/alexa-top-sites/`.
[7] Repository of IDN Practices: `https://www.icann.org/resources/pages/idn-guidelines-2003-06-20-en`.
[8] Repository of IDN Practices: `https://www.iana.org/domains/idn-tables`.
[9] Costello, A. M.: Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA), RFC 3492 (2003).
[10] mozilla: IDN Display Algorithm, `https://www.chromium.org/developers/design-documents/idn-in-google-chrome`.
[11] The Chromium Projects: IDN in Google Chrome, `https://www.chromium.org/developers/design-documents/idn-in-google-chrome`.
[12] The Unicode Consortium confusables.txt: `https://unicode.org/Public/security/11.0.0/confusables.txt`.
[13] 上綱秀治：新旧字体表，`http://www.asahi-net.or.jp/~ax2s-kmtn/ref/old_chara.html`.
[14] 社団法人 日本ネットワークインフォメーションセンター：汎用 JP ドメイン名登録等に関する技術細則，`https://jprs.jp/doc/rule/saisoku-1-wideusejp.html`.
[15] Sawabe, Y., Chiba, D., Akiyama, M. and Goto, S.: Detecting Homograph IDNs Using OCR, *Proc. of the 15th APAN Research Workshop 2018*, pp. 56–64 (2018).