氏名	NGUYEN SON HOANG QUOC	
学位(専攻分野)	博士(情報学)	
学位記番号	総研大甲第 1798 号	
学位授与の日付	平成27年9月28日	
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当	
学位論文題目	ANONYMIZING PRIVATE PHRASES AND DETECTING DISCLOSURE IN ONLINE SOCIAL NETWORKS	
論文審査委員	主査	教授　越前　功
		教授　曽根原　登
		准教授　宮尾　祐介
		准教授　岡田　仁志
		特任教授　山田　茂樹　国立情報学研究所
		教授　小舘　亮之　津田塾大学

(別紙様式 2)
(Separate Form 2)

論文内容の要旨
Summary of thesis contents

ANONYMIZING PRIVATE PHRASES AND DETECTING DISCLOSURE IN ONLINE SOCIAL NETWORKS

Online social networks (OSNs) have become an important part of modern life, and many people use one or more OSNs every day. However, the ease of collecting personal information from OSN messages can make users feel insecure when they access an OSN. Phrases containing personal information, i.e., “private phrases,” should thus be identified and anonymized before they are posted on OSNs. Furthermore, if messages containing personal information that are posted only for friends to see are disclosed, the discloser should be identifiable.

Most previous research on identifying private phrases has focused on comparing candidate phrases with predefined phrases (such as personal names, locations, or diseases). However, attackers easily recognize and replace the phrases with similar ones (e.g., synonyms, generalizations). Such replacements can be identified by using a co-occurrence metric. In this case, non-private phrases, such as HIV in a non-private message “The human immunodeficiency virus (HIV) is a lentivirus (a subgroup of retrovirus) that causes the acquired immunodeficiency syndrome (AIDS),” might also be identified.

We have developed an algorithm for determining whether an OSN message is a private message or a non-private one. The F-score was 92% for 3000 messages, showing that it works well for identifying private messages. The algorithm identifies private messages on the basis of word frequency, so it is not suitable for private messages containing locational phrases. Consequently, we have developed a rule-based algorithm that overcomes this problem by using text semantics. It correctly identified 84.95% of 2917 locational messages, significantly higher than a machine learning method using word frequency and its three extensions (highest accuracy=81.93%). Phrases that are identified as private might be released for only friends to see. For example, users sometimes share messages containing such information as e-mail addresses, hometowns, and locations. This information should be anonymized to avoid spam or advertising. One approach to such anonymization is to remove or replace the private phrases with generalizations, but this makes the text sound unnatural.

We have developed a way to improve the naturalness of generalized phrases. We have also developed a metric for quantifying information loss due to generalization so that anonymized messages can be distributed to different groups of friends with appropriate levels of privacy.

Time-related information in texts posted on-line is one type of private information targeted by attackers. This is one reason that sharing information online can be risky. We have developed an algorithm for creating anonymous fingerprints for temporal

phrases that covers most potential cases of private information disclosure. The fingerprints not only anonymize time-related information but also can be used to identify a person who has disclosed information about the user. In an experiment with 16,647 different temporal phrases extracted from about 16 million tweets, the average number of fingerprints created for an OSN message was 526.05. This is significantly better than the 409.58 fingerprints created by a state-of-the-art algorithm. Fingerprints are quantified using a modified normalized certainty penalty metric to ensure that an appropriate level of information anonymity is used for each friend of the user.

The algorithm we developed to anonymize time-related private information removes the temporal phrases when doing so will not change the natural meaning of the message. The temporal phrases are detected by using machine-learned patterns, which are represented by a subtree of the sentence parsing tree. The temporal phrases in the parsing tree are distinguished from other parts of the tree by using temporal taggers integrated into the algorithm. In an experiment with 4008 sentences posted on a social network, 84.53% of them were anonymized without changing their intended meaning. This is significantly better than the 72.88% of the best previous temporal phrase detection algorithm. Of the learned patterns, the top ten most common ones were used to detect 87.78% of the temporal phrases. This means that only some of the most common patterns can be used to anonymize temporal phrases in most messages to be posted on an OSN.

As mentioned above, private messages could be revealed by a user's friends or even by the user, either unintentionally or intentionally. One approach to overcoming this problem is to create fingerprints by using linguistic steganography algorithms. Unfortunately, such algorithms create an insufficient number of fingerprints to cover all of a user's friends.

The algorithm we developed generates a sufficient number of fingerprints by using various combinations of synonymizations and generalizations. It creates, on average, 140.91 fingerprints per message. This is significantly higher than the 21.29 fingerprints created by the best fingerprinting algorithm using synonymization. In addition, attackers often modify their fingerprints into paraphrased ones. The similarity matching (SimMat) metric we developed for detecting paraphrases is based on matching identical phrases and similar words and quantifying the minor words. Evaluation using about 5800 paraphrase pairs taken from a widely used paraphrase corpus created by Microsoft Corporation demonstrated the effectiveness of the SimMat metric. It achieved the highest paraphrase detection accuracy (77.6%) when it was combined with eight standard machine translation metrics. This accuracy is slightly better than the 77.4% rate achieved with the current state-of-the-art method for paraphrase detection.

We have developed a realistic system that controls the disclosure of both private and non-private messages on Facebook, the largest OSN. It automatically identifies private messages after a user composes them using the developed system. The system then recommends anonymous fingerprints for a user's friends on the basis of private phrases in the private messages. This system also detects the disclosure of fingerprinted information and the person who disclosed it.

The system not only efficiently controls the disclosure of private information in OSN messages but also improves security in other sensitive fields (e.g., health, military, and politics). Furthermore, our proposed algorithms have been proven to be common types of attack (such as paraphrasing and generalizing). Although they are poor at identifying private information in a single message, such information can be inferred from various messages (past OSN messages, blogs, web pages, etc.).

Future work includes enhancing the algorithms to enable them to identify private information by inferring it from various messages. In the next stage, we will focus on improving the naturalness and semantic coherence of anonymous fingerprints. In addition, the algorithms will be modified to enable them to anonymize private objects (such as human faces, vehicle license numbers, and home addresses) in digital media (audio, image, video, etc.).

(別紙様式 3)
(Separate Form 3)

博士論文の審査結果の要旨
Summary of the results of the doctoral thesis screening

出願者，Nguyen Son Hoang Quoc 氏は，「Anonymizing Private Phrases and Detecting Disclosure in Online Social Networks」と題する論文を提出し，この論文およびその内容に基づく研究発表に基づき博士論文の審査が行われた．本論文は，Online Social Networks (OSNs)上などで共有される自然言語によるメッセージの匿名化手法と，当該メッセージの漏えい者を特定する手法の確立を目的としている．具体的には，メッセージから投稿者や特定のユーザに関する private phrase をルールベースにより抽出し，抽出した private phrase を OSNs 上のユーザ属性（例 families, friends, or acquaintances）に従って，異なる一般化レベルで匿名化するとともに，匿名化の多義性（同一の匿名性を確保するために多様な匿名化プロセスが存在する）に基づいて，ユーザの識別情報と匿名化プロセスを紐づけることで，メッセージの匿名化と同時にユーザの識別情報をメッセージに埋め込み検知する匿名化フィンガープリント手法を提案した．

本論文は 6 章から構成される．まず第 1 章では，OSNs 上で生じるプライバシー侵害を概観するとともに，本研究の目的である OSNs 上で共有されるメッセージの匿名化手法とメッセージ漏えい者の特定手法の必要性について述べている．また，本研究の目的を実現するための提案手法の 3 つの技術課題（メッセージ内の private phrase の抽出，private phrase の匿名化，メッセージ漏えい者の特定）について述べるとともに，これらの技術課題の克服により OSNs 上で実装可能なシステム提案について述べている．第 2 章では，提案手法の 3 つの技術課題それぞれについて関連研究を分析・比較し，提案手法の新規性および有用性について述べている．

第 3 章では，提案手法を実現するための最初の技術課題である，OSNs 上で共有されるメッセージからの private phrase の抽出方法について述べている．具体的には，メッセージから人物,場所,組織などの phrase を candidate phrase として抽出した後に，candidate phrase 間の関係をルールベースにより分析することで，当該メッセージが普遍的な真理や一般的事実などを述べた non-private message か，特定の個人に関する private message か判別し，private message であれば，当該メッセージから抽出した candidate phrase を private phrase とする．Tweet2011 dataset を用いた評価実験では，従来から存在する 4 つのアプローチと抽出精度を比較し，提案方法が最も高い精度であることを示した．

第 4 章では，提案手法を実現するための 2 番目の技術課題である，OSNs 上で共有されるメッセージから抽出した private phrase の匿名化方法について述べている．具体的には，OSNs 上で投稿者のメッセージを他ユーザと共有する際に，他ユーザのユーザ属性（例 families, friends, or acquaintances）に応じて，異なる一般化レベル（例 Tokyo から Japan/Asia）で private phrase を匿名化する．また，匿名化の多義性（同一の匿名性を確保するために多様な匿名化プロセスが存在する）に基づいて，ユーザの識別情報と匿名化プロセスを紐づけることで，メッセージの匿名化と同時にユーザの識別情報をメッセージに埋め込む匿名化フィンガープリントの提案について述べている．

第 5 章では，提案手法を実現するための 3 番目の技術課題である，投稿者のメッセージを他ユーザが第三者に漏えいした際の当該メッセージおよび漏えい者の検知方法について述べている．具体的には，投稿者のメッセージが他ユーザから第三者を経由して伝達される際にメッセージの言語表現に言い換え(paraphrase)が生じるが，言い換えが行われたメッセージから元のメッセージを特定する paraphrase detection の提案について述べてい

る．漏えい者の特定については，4 章で述べた方法に基づいて，特定されたメッセージの匿名化プロセスを解析することで漏えいしたユーザを特定する．

第 6 章では，結論として，本論文の貢献についてまとめ，同分野における今後の研究課題について述べている．

本研究の成果は，査読付学術雑誌論文 1 篇，査読付国際会議 6 篇，その他学会発表 4 篇（いずれも出願者が主著）として発表されており，研究内容が国内外で認められていることを示している．以上の学術的貢献を総合的に判断して，本博士論文は学位を授与するに値すると判断した．