LETTER *Special Section on Information and Communication System Security*

Examining Privacy Leakage from Online Used Markets in Korea

Hyunsu MUN[†a)], ***Nonmember and*** **Youngseok LEE**[†b)], ***Member***

SUMMARY Online used markets such as eBay, Yahoo Auction, and Craigslist have been popular due to the web services. Compared to the shopping mall websites like eBay or Yahoo Auction, web community-style used markets often expose the private information of sellers. In Korea, the most popular online used market is a website called "Joonggonara" with more than 13 million users, and it uses an informal posting format that does not protect the users' privacy identifiable information. In this work, we examine the privacy leakage from the online used markets in Korea, and show that 45.9% and 74.0% of sample data expose cellular phone numbers and email addresses, respectively. In addition, we demonstrate that the private information can be maliciously exploited to identify a subscriber of the social network service.

key words: *privacy, online used market, community, web, crawl, Facebook*

1. Introduction

Privacy has been an important issue on the web, social network service, and commercial sites [1]–[6]. Third-party scripts or programs on the website related with advertisements track personal information with cookies [1]. Typically, privacy leakage comes from insecure data store, communication, or software. Even secure protocols can be cracked to reveal private information [5]. Recently, it is reported that while using shopping websites [4], hospital reservation service [5], or social network services [2], [3], [6], people sometimes leave privacy footprints which can be maliciously used to track users.

Online used markets or flea markets have been popular due to the development of web services. eBay[*], Yahoo Auction[**], and Craigslist[***] are the representative online used markets. In eBay or Yahoo Auction, people buy used products in auction or in a regular procedure of the commercial shopping website such as Amazon, Rakuten, or Taobao. On the other hand, Craigslist is a flea market website where a seller posts an informal message about the product for sales with the contact information.

In general, the online used market websites show the seller's contact information because customers need to ask questions about product or delivery, or to leave the review. These online shopping mall websites provide the formal communication tools such as message boxes or dedicated emails in order to deal with the customers' questions and feedback. Therefore, it is not necessary for sellers to put their private contact information on the sales web page. On the other hand, in flea markets such as Craigslist, sellers tend to post their messages with their private contact information such as email addresses or cellular phone numbers, because the flea market websites usually use a simple text form to explain information about the product status and the price as well as the contact information. In Craigslist, a seller and a customer exchange the product with the money according to their agreement on price, delivery, and payment method through various ways of communication media such as emails, text messages, or phone calls. Because of the recent privacy rule of Craigslist, the private information of personal sellers is scarcely observed, whereas professional sellers often leave their office addresses and phone numbers.

In Korea, the most popular online used market is a web community called "Joonggonara"[****] which has more than 13 million users and 150,000 messages for sales per day. In the Joonggonara website, sellers post their messages explaining information about the product status, price, or delivery method. A message posted to Joonggonara is divided into header and body areas as shown in Fig. 1. A header of the message consists of product category, title, phone number, and email address fields. A body has two sub-parts: the first part is a table form of sales information including seller ID, email address, location, product name, price, and delivery option; the second part is the text space where sellers can write their own comments. Phone numbers and email addresses are explicitly exposed to users who visit the message.

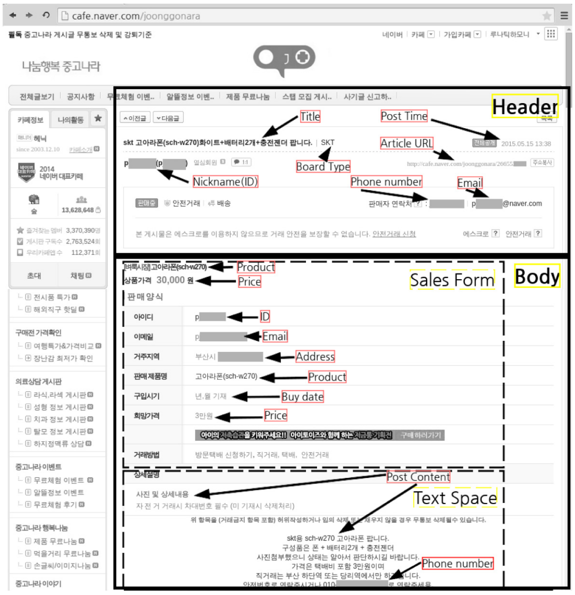
In order to protect the user privacy, the Joonggonara website has a simple policy that hides phone numbers and email addresses on the header either after one month since the posted message was modified or when the seller changes the status of the transaction as "sold out". However, sellers often write their phone numbers or email addresses within the message body, which will be exposed to the public until the posted message is deleted by the seller. Popular web communities in Korea such as "Ppomppu" (a community sharing shopping information[*****]) and "Ruliweb" (a com-

Manuscript received June 2, 2015.
Manuscript revised October 8, 2015.
Manuscript publicized January 13, 2016.

[†]The authors are with the Dept. of Computer Engineering, Chungnam National University, Korea.
a) E-mail: munhyunsu@cnu.ac.kr
b) E-mail: lee@cnu.ac.kr (Corresponding author)

DOI: 10.1587/transinf.2015ICL0003

[*] http://www.ebay.com
[**] http://auction.yahoo.co.jp
[***] http://www.craigslist.org
[****] http://cafe.naver.com/joonggonara
[*****] http://www.ppomppu.co.kr



skt 고아라폰(sch-w270)화이트+배터리2개+충전젠더 팝니다. | SKT   전체공개 2015.05.15 13:38

p[redacted](p[redacted]) 열심회원 1:1   http://cafe.naver.com/joonggonara/26655[redacted] 주소복사

판매중 안전거래 | 배송   판매자 연락처 : [redacted] | p[redacted]@naver.com

본 게시물은 에스크로를 이용하지 않으므로 거래 안전을 보장할 수 없습니다. 안전거래 신청   에스크로 ? 안전거래 ?

[벼룩시장]고아라폰(sch-w270)

상품가격 30,000 원

판 매 양 식

아이디	p[redacted]
이메일	p[redacted]
거주지역	부산시 [redacted]
판매제품명	고아라폰(sch-w270)
구입시기	년,월 기재
희망가격	3만원
	아이의 저축습관을 키워주세요!! 아이토이즈와 함께 하는 저금통 기획전 구매하러가기
거래방법	방문택배 신청하기, 직거래, 택배, 안전거래

상세설명

사진 및 상세내용

자 전 거 거래시 차대번호 필수 (미 기재시 삭제처리)

위 항목을 (거래금지 항목 포함) 허위작성하거나 임의 삭제 또는 채우지 않을 경우 무통보 삭제될수 있습니다.

skt용 sch-w270 고아라폰 팝니다.
구성품은 폰 + 배터리2개 + 충전젠더
사진첨부했으니 상태는 알아서 판단하시길 바랍니다.
가격은 택배비 포함 3만원이며
직거래는 부산 하단역 또는 당리역에서만 하겠습니다.
안전번호로 연락주시거나 010-[redacted]로 연락주세요

Fig. 1 The most popular online used market in Korea, “Joonggonara” and its posting format.

munity sharing digital game information[†]) also have their own used markets whose sales forms are similar with that of Joonggonara. Thus, online used markets including the personal contact information are vulnerable to privacy leakage and related threats.

In this work, we present how privacy leaks from online used markets in Korea by investigating representative used markets of web communities. From experiments, we have found that 45.9% and 74.0% of messages for sales in Joonggonara expose privacy identifiable information such as phone numbers and email addresses, respectively. We have also shown that the phone numbers and email addresses can be exploited to identify a person of the social network service, Facebook. From the collected phone numbers and email addresses, we have identified 46.8% of online used market users at the Facebook.

2. Privacy Data Collection and Analysis Method

We have designed and implemented a simple privacy data collection and analysis system with functions of data collection, sales information extraction, and privacy analysis as shown in Fig. 2. We implemented a web community crawler that gathers web pages from online markets based on Python Scrapy[††]. After collecting web pages, we have extracted the sales information about the product name, price, seller ID, email address, and phone number from the message header and the body with sales form and text space parts. As the text space part is written in a natural language, we have built a parser using Beautiful Soup[†††] to find the privacy information elements such as phone number, email address, and location. We used a simple string pattern matching algorithm for each privacy information element. After the data cleansing phase, we loaded the sales transaction data with the privacy information elements into tables of MySQL database. By integrating R with MySQL database, we analyzed results with plots.

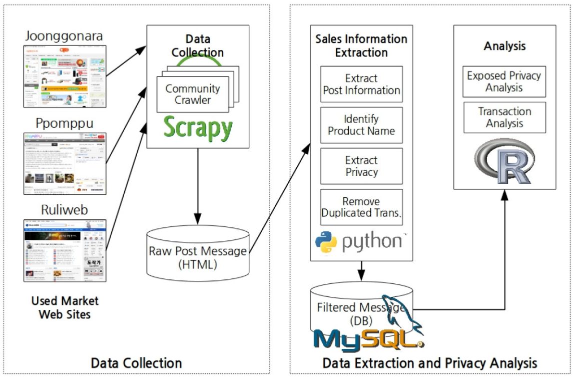


Fig. 2 The overview of privacy data collection and analysis system.

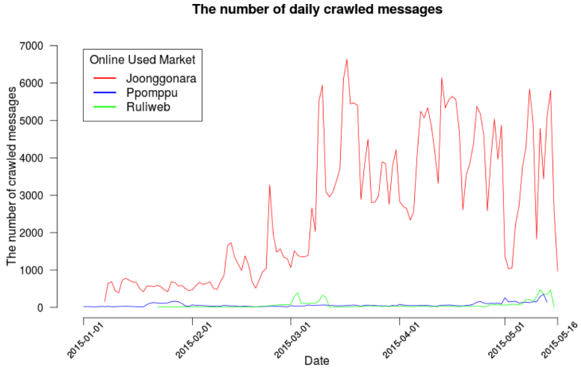


Fig. 3 The number of daily crawled messages from online used markets in Korea.

3. Analysis of Privacy Footprints

3.1 Data

We collected the sales messages from three representative online used markets of web communities (Joonggonara, Ppomppu, and Ruliweb). In accordance with the crawling policy of each website, we gathered sample data from the website to avoid the traffic overload to the web server. As the number of sales messages posted to the Joonggonara website is too large, we selected a cellular phone category which is one of the popular item in the used market, and collected a sample of messages for sales every day. After starting with a small amount of messages from January 2015, we gradually increased the sample space, as shown in Fig. 3.

The summary of the crawled data is explained in Table 1. The total number of messages in Joonggonara is 331,955 and outnumbers two websites. We consider the privacy identifiable information as id, nickname, phone number, email address, location, and IP address. While id, phone

[†] http://www.ruliweb.com
[††] http://scrapy.org
[†††] http://www.crummy.com/software/BeautifulSoup

Table 1 The summary of collected data from online used markets in Korea: total sales message count, collection period, privacy information.

Market	Message	Period	Privacy data
Joonggonara	331,955	2015-01-07 ˜ 2015-05-16	id, nickname, phone, email, location
Ppomppu	8,551	2014-12-30 ˜ 2015-05-13	nickname, phone
Ruliweb	7,077	2015-01-22 ˜ 2015-05-15	nickname, phone, email, location, IP

Table 2 The privacy leakage results in online used markets (%).

Market	Phone Number	Email	Location	IP
Joonggonara	45.9	74.0	16.4	-
Ppomppu	31.0	-	-	-
Ruliweb	96.8	0.0	100.0	100.0

number, and email address are commonly found in these online used markets, location and IP addresses are often observed. In Ruliweb, the IP address is exposed so that we can estimate the geolocation with the IP address.

3.2 Privacy Leakage Results

The structure of a message posted to Joonggonara is divided into header and body areas. The Joonggonara website removes the phone number and email address on the header of the message either after one month since the message was modified or the status of the message is changed to "sold out". Though the Joonggonara website deletes the private information, this policy is not applied to the whole message body. Therefore, the private information on the message body will be exposed until the message is removed by the seller.

Table 2 shows the result of privacy leakage from the whole sales messages in three online used markets. In Joonggonara, the phone number leakage ratio is 45.9%. In addition to phone numbers, we can observe email addresses from 74.0% of messages. Sellers often put the specific location name around home or office for in-person sales. The location information was found by 16.4% of messages in Joonggonara. In the Ppomppu online website, we obtained phone numbers by 31.0%, but the email addresses are not exposed. In the Ruliweb market, it is interesting that we can find the IP address of the message by 100%. The IP addresses are automatically collected by the Ruliweb server and inserted in the message. According to the policy of the Ruliweb online used market, it is recommended that users check whether the IP address in the message belongs to the location within Korea, or not. Thus, Ruliweb always marks the location information with IP geolocation database. In addition, phone numbers are frequently found in Ruliweb by 96.8%.

As the privacy policy of Joonggonara is applied to the header of the posted messages, we can still observe private information on the message body.

In Table 3, we show the percentage of phone numbers and email addresses by the header or the body of messages.

Table 3 The privacy (phone number, email) leakage by two areas of a message in Joonggonara.

Area	Phone Number (%)	Email (%)
Header	29.1	73.5
Body	32.4	22.9
Header or Body	45.9	74.0

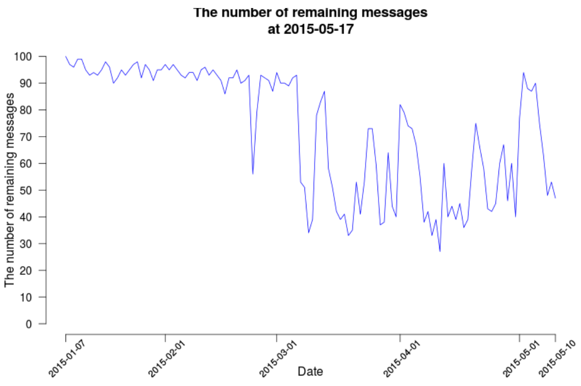


Fig. 4 Messages among 100 daily samples that remain at 2015.5.17.

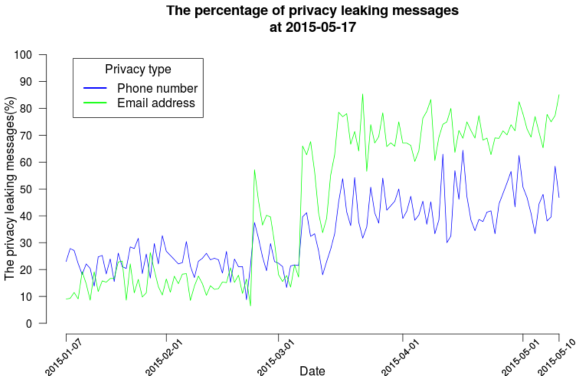


Fig. 5 Privacy footprint decay of the remaining messages at 2015.5.17.

29.1% of phone numbers are observed in the header of the messages and 32.4% in the body. The percentage of phone numbers exposed on either the head or the body area is 45.9%. Email addresses on the message body are less observed (22.9%) compared with the phone numbers, but they appear more often on the message header (73.5%).

Sellers can decide whether to delete their posted messages or not, because messages with the mark "sold out" can be useful for enhancing the credit of the seller. Therefore, we examine how the number of posted messages and their privacy identifiable information vary over the time. We randomly selected 100 messages per day from three online markets and probed each message at a specific time (2015.5.17) whether the message still remains or not. From Fig. 4, we can know that a lot of messages are visible since the messages were first posted. It is interesting that 95 messages posted on January are not deleted after four months.

Next, we evaluated the privacy leakage of the remaining messages among 100 samples. Figure 5 illustrates the percentage of phone numbers and email addresses of the re-

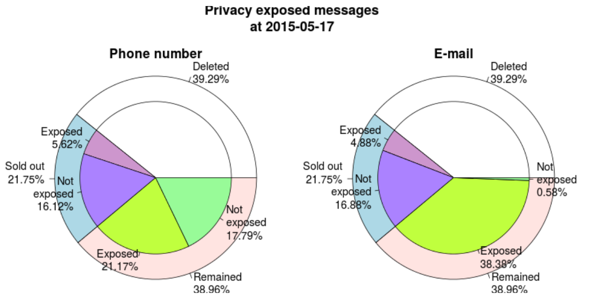
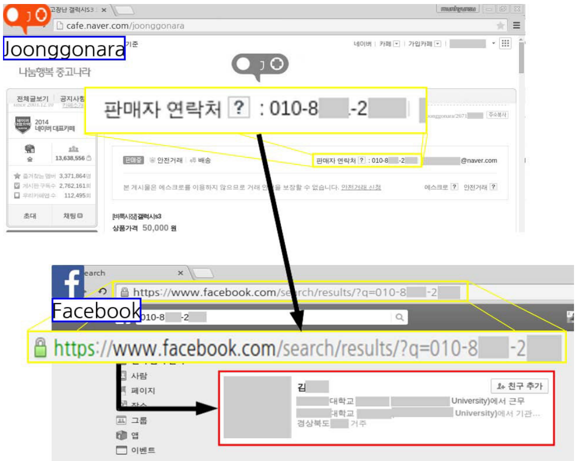


Fig. 6 Classification of messages with or without privacy information from 100 daily samples during 2015.4.17.˜2015.5.10. at 2015.5.17.


판매자 연락처 ? : 010-8[redacted]-2[redacted]

판매자 연락처 ? : 010-8[redacted]-2[redacted] [redacted]@naver.com

본 게시물은 에스크로를 이용하지 않으므로 거래 안전을 보장할 수 없습니다. 안전거래 신청

[버룩시장] 갤럭시s3

상품가격 50,000 원

https://www.facebook.com/search/results/?q=010-8[redacted]-2[redacted]

김[redacted]

[redacted]대학교 [redacted] [redacted] University)에서 근무

[redacted]대학교 [redacted] [redacted] University)에서 기관...

경상북도[redacted] 거주

Fig. 7 Identifying the subscriber of Facebook with the phone number found from the online used market.

maining messages. As shown in Fig. 5, the privacy footprint gradually decreases as time goes on. However, phone numbers do not quickly disappear compared with email addresses, because phone numbers are often observed in the text space of the message body. 23.3% and 23.4% of the remaining messages in January and February 2015 expose phone numbers as of May 2015 (35.4% in March, 43.9% in April). 14.9% and 20.3% of the remaining messages in January and February 2015 expose email addresses as of May 2015 (55.7% in March, 70.9% in April).

Sellers are aware of privacy exposure and take steps to protect their cellular phone numbers or emails. After the sale transaction, sellers often delete or mark their posts as “sold out” which removes the privacy in the header of the post message. In Fig. 6, we examined how many messages are protected by sellers or the web server policy that hides the header information of the seller after one month. 61.04% of messages were taken steps to protect privacy information by sellers (“Sold out” and “Deleted”). However, 5.62% and 4.88% of sold out messages still expose phone numbers and email addresses.

In order to show the vulnerability of privacy leakage, we have examined whether we can identify a subscriber of Facebook with the phone number obtained from the online used markets, or not. In accordance with the privacy policy of Facebook, we have searched a small number of sample messages. Given with 1,748 sample messages with phone numbers, we can identify 819 subscribers of Facebook, which is 46.8%. Figure 7 shows how we probed the subscriber of Facebook with the phone number found from the online used market.

4. Conclusion

In this work, we have examined the privacy leakage of online used markets in Korea. Through the experiments, we have found that representative online used markets in Korea are vulnerable to expose the privacy. In particular, we revealed that 45.9% of messages in Joonggonara expose phone numbers and 74.0% of messages contain email addresses. We have also known that the location and IP addresses are observed from a website called Ruliweb.

As the current privacy policy of online used markets in Korea is too primitive, it is essential to strengthen the privacy protection method. The most secure method will be to prohibit the private information within the posted message and to provide randomized identifiers and website-specific communication tools like Craigslist. On the other hand, the web server can deploy a more strict privacy protection policy that detects and protects the string pattern related with the privacy in the sales message.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology [NRF-2013R1A1A2007326].

References

[1] D. Malandrino and V. Scarano, “Privacy leakage on the Web: Diffusion and countermeasures,” Computer Networks, vol.57, no.14, pp.2833–2855, Oct. 2013.

[2] R. Dey, Y. Ding, and K.W. Ross, “Profiling high-school students with facebook: how online privacy laws can actually increase minors’ risk,” Proceedings of the 2013 ACM conference on Internet measurement conference (IMC ’13), pp.405–416, 2013.

[3] R. Dey, C. Tang, K.W. Ross, and N. Saxena. “Estimating age privacy leakage in online social networks,” Proceedings of the IEEE INFOCOM 2012, Orlando, FL, USA, pp.2836–2840, 2012.

[4] T. Minkus and K.W. Ross, “I Know What You’re Buying: Privacy Breaches on eBay,” Lecture Notes in Computer Science vol.8555, pp.164–183, 2014.

[5] B. Miller, L. Huang, A.D. Joseph, and J.D. Tygar, “I Know Why You Went to the Clinic: Risks and Realization of HTTPS Traffic Analysis,” Lecture Notes in Computer Science, vol.8555, pp.143–163, 2014.

[6] C. Fu, Z. Shaobin, S. Guangjun, and G. Mengyuan, “Crowdsourcing Leakage of Personally Identifiable Information via Sina Microblog,” Lecture Notes in Computer Science vol.8662, pp.262–271, 2014.