2015

Attack-Resistant Digital Reputation and Privacy Assessment in Social Media

Yongbo Zeng
University of Rhode Island, bo.li.zeng@gmail.com

Follow this and additional works at: https://digitalcommons.uri.edu/oa_diss

Recommended Citation

Zeng, Yongbo, "Attack-Resistant Digital Reputation and Privacy Assessment in Social Media" (2015). *Open Access Dissertations*. Paper 409.
https://digitalcommons.uri.edu/oa_diss/409

This Dissertation is brought to you for free and open access by DigitalCommons@URI. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons-group@uri.edu.

ATTACK-RESISTANT DIGITAL REPUTATION AND PRIVACY
ASSESSMENT IN SOCIAL MEDIA

BY
YONGBO ZENG

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
ELECTRICAL, COMPUTER & BIOMEDICAL ENGINEERING

UNIVERSITY OF RHODE ISLAND

2015

DOCTOR OF PHILOSOPHY DISSERTATION
OF
YONGBO ZENG

APPROVED:

Dissertation Committee:

Major Professor Yan Sun

Tao Wei

Lisa DiPippo

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2015

ABSTRACT

Web 2.0 has been growing rapidly in the past decade, and leading to surging popularity of online social media. There are over 2.1 billion people that are using social media, which is 28% of the global population. Social media has become one of the most complex computing and communication systems in the planet. Social media attracts large amount of people to create, share and exchange information, interests, ideas, pictures, videos, and etc. in the virtual communities. In social media, people can interact with acquaintances and strangers, and thus privacy and security should be considered seriously.

From the *privacy* perspective, one of the severe type of privacy breach is related to online social networks, such as Facebook, Linkedin, Google+, and Twitter. Online social network users are often not aware of the size and the nature of the audience viewing their profiles, and therefore they may reveal more information than what is appropriate to be viewed publicly. Due to the lack of privacy awareness, online social network users can suffer a number of privacy related threats. In this dissertation, a quantitative online social network privacy risk analysis framework – TAPE is proposed. Inspired by the reliability analysis of a wireless sensor network, the binary decision diagram tool is employed to calculate online social network privacy level. The privacy awareness and privacy trust metrics are proposed to evaluate online social network users' intention of privacy protection. To our best knowledge, TAPE framework is the first work that take both privacy awareness and privacy trust into consideration. Based on the TAPE framework, we also propose an unfriending strategy in terms of privacy protection, which outperforms other existing unfriending strategies. The detail of this framework is introduced in **Chapter 2**.

From the *security* perspective, online product/service review system is one

of the most vulnerable systems in social media. Since there are enormous profits of online markets and the customers' purchasing decision is relying on the product/service review, it is highly possible that firms and retailers at the online marketplace may create fake reviews to mislead customers. In this dissertation, a novel angle of fake review detection is introduced, which is called Equal Rating Opportunity (ERO) principle. Based on ERO principle, ERO analysis is proposed. ERO analysis can be implemented with limited cost. It is a new direction of fake review detection. Based on real data testing, ERO analysis is able detect new perspectives of fake review, which cannot be detected by other approaches, while giving a relatively low false alarm rate. The ERO principle and ERO analysis is presented in **Chapter 3**.

ACKNOWLEDGMENTS

There are so many people to thank for helping me during my PhD study at URI. So many have made my stay here productive and pleasant. I will try to cover all the bases without long winded words.

Foremost, I would like to thank my advisor and mentor, Dr. Yan Lindsay Sun, for her guidance, encouragement and inspiration over the past four years. Dr. Sun introduced me to the wonderland of research. She helped me thrive in both academic and social abilities. She taught me many skills in writing and presenting, which I believe will benefit in all my life. Whenever I discussed with her my problems and ideas, she was always a good listener, and she was always able to give me constructive suggestions. Her support was essential to my success, and it paved the way to this dissertation.

I would like to sincerely thank Dr. Tao Wei, Dr. Haibo He, Dr. Lisa DiPippo and Dr. Li Wu for serving on my dissertation committee. I really enjoyed the time talking and discussing with them. Thank them all for their support and constructive suggestions to improve the quality of this dissertation.

I would like to thank my collaborators, Dr. Liudong Xing, Dr. Vinod Vokkarane and Dr. Chaonan Wang. We worked on the online social network privacy project. I got a lot of helps from them to understand the mathematical model of reliability graph and to build the simulation system. They also gave me advices on manuscript writing and revision.

I would also like to thank many professors and staffs in ECBE department. I learn from many of them via different courses, which equipped me with knowledge to tackle problems in this dissertation and in my future career. The department staff were also very helpful, especially Meredith Leach Sanders. Meredith helped me to deal with paperwork and showed me concrete guidance of academic affairs.

Special thanks to colleagues and alumnus as well as friends in the Network Security and Trust Laboratory Dr. Yihai Zhu, Dr. Yuhong Liu, Dr. Wenkai Wang, and Dr. Yafei Yang for their support and help in my PhD study.

Special thanks to my friends, Daxian Yun, Dr. Zhen Ni, Dr. Quan Ding, Zhen Chen, Yazan N. Rawashdeh, Jun Yan, Yufei Tang, Jing Yang and many others, for their helps when I was pursuing my PhD.

Finally and most importantly, none of this would have been possible without the love, patience and support of my family. I would like to express my heartfelt thanks to my family. My parents raised and educated me with their unconditional love. My wife, Dr. Li Gao, always accompanied and encouraged me with her eternal and unchanging love. I would like to express my thanks to my father-in-law, mother-in-law and brother-in-law. I would also like to thank other family members for their supports in my life.

PREFACE

This dissertation is organized in the manuscript format. Particularly, there are three chapters. The introduction is given in Chapter 1, followed by two manuscripts discussed in Chapter 2 and Chapter 3. A brief introduction of the manuscripts are as follows.

- Manuscript 1 in Chapter 2:

Yongbo Zeng, Yan (Lindsay) Sun, Liudong Xing, and Vinod Vokkarane, “Online Social Networks Privacy Study Through TAPE Framework”, IEEE Journal of Selected Topics in Signal Processing, 2015, in press

- Manuscript 2 in Chapter 3:

Yongbo Zeng, Yihai Zhu, and Yan (Lindsay) Sun, “Equal Rating Opportunity Analysis for Detecting Review Manipulation”, in preparing for submission to IEEE Transactions on Information Forensics and Security. An earlier and shorter version is published in the IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP 2015)

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGMENTS	iv
PREFACE	vi
TABLE OF CONTENTS	vii
LIST OF TABLES	xi
LIST OF FIGURES	xii
CHAPTER	
1 Introduction	1
1.1 Online Social Network Privacy	2
1.1.1 Privacy Threats	2
1.1.2 Privacy Protection	3
1.1.3 Quantitative Privacy Risk Analysis	4
1.2 Security of Online Review System	6
1.2.1 Online Review System	6
1.2.2 Fake Review	7
1.2.3 Online Review System Protection	8
1.3 Summary	9
List of References	10
2 “Online Social Networks Privacy Study Through TAPE Framework”	15
2.1 Abstract	16

	Page
2.2 Introduction	16
2.3 Related Work	20
2.4 Trust-aware Privacy Evaluation Framework	22
2.4.1 Acronyms	23
2.4.2 Notations	23
2.4.3 Online Social Network Privacy	25
2.4.4 Privacy Risk and Related Concepts	27
2.4.5 Toward Privacy Leakage Probability Estimation	30
2.4.6 Privacy Analysis and Reliability Analysis	31
2.4.7 Summary	35
2.5 Information Spreading Probability Algorithms	35
2.5.1 Node Information Spreading Probability (NISP)	36
2.5.2 Link Information Spreading Probability (LISP)	44
2.6 Privacy Assessment and Privacy Improvement through TAPE	45
2.6.1 Privacy Assessment	45
2.6.2 Privacy Improvement Strategies	45
2.7 Experiment Results and Discussion	48
2.7.1 Case Study	48
2.7.2 Datasets	50
2.7.3 Privacy Risk	51
2.7.4 The impact of PA and PT	54
2.7.5 Verification of TAPE Calculation	55
2.7.6 Sensitivity Analysis and Unfriending Strategy	57

	Page
2.7.7 Discussion	59
2.8 Conclusion and Future Work	60
List of References	60
3 “Equal Rating Opportunity Analysis for Detecting Review Manipulation”	65
3.1 Abstract	66
3.2 Introduction	66
3.3 Related Work	70
3.4 ERO Principle and ERO Analysis	72
3.4.1 Consistency Analysis	72
3.4.2 ERO Principle	74
3.4.3 ERO Feature	75
3.4.4 ERO Analysis	76
3.4.5 Case Study	77
3.5 ERO Feature Selection	79
3.5.1 Review Feature Categorization	79
3.5.2 Feature Comparison	81
3.6 Experiment and Results	84
3.6.1 Performance Evaluation	84
3.6.2 Dataset	85
3.6.3 Victim Product Detection Comparison	86
3.6.4 Expert Reviews	88
3.6.5 Discussion	89
3.7 Conclusion	90

List of References	91
------------------------------	----

LIST OF TABLES

Table		Page
1	Concepts mapping	31
2	Desirable properties of PA algorithms	37
3	An example: privacy setting statistics for birthday	39
4	Case study: Rank PA vs. IRT PA	47
5	Datasets summary	51
6	Impact of LISP value when comparing TAPE and the topology-based method	55
7	Statistic summary of unfriending strategies	58
8	Review features	81
9	Review features summary	84
10	Review feature crawled from Amazon	85
11	Summary of dataset	86
12	Experiment parameters	87
13	Expert review	88

LIST OF FIGURES

Figure		Page
1	Online social network of Example 1.	25
2	Core structure of the TAPE framework	28
3	Similarity between WSN and OSN.	32
4	An OSN information diffusion example.	33
5	BDD graph of the example in Fig 4	34
6	Privacy Trust model	41
7	Node ISP Calculation Diagram	44
8	Privacy risk vs. PIO UD distance.	52
9	Histogram of proportional difference	53
10	TAPE vs Monte-Carlo simulation	56
11	Privacy improvement of unfriending	57
12	ERO principle examples	78
13	Review features	82
14	Fake review detection comparison	88

CHAPTER 1

Introduction

In the past decade, there was a dramatic rise in popularity of online social media, such as micro-blogs, online social networks, online product/service review systems, crowdfunding, video sharing, photo sharing, business network, social gaming, and etc. It is reported that by January 2015, there are over 3 billion active Internet users, around 2.1 billion people have active social media accounts, and the population of social media is still growing [1]. Social media has become one of the most complex computing and communication system on the planet. Just like any modern computing and communication systems, security and privacy are the primary concerns.

From the *privacy* perspective, one of the severe types of privacy breach is related to online social networks (OSNs). Facebook, Linkedin, Google+, and Twitter are some of the most visited OSNs. According to the report by Kemp, Facebook has 1.4 billion users and 4.5 billion daily likes, Twitter has 284 million users and 500 million daily tweets, and Google+ has 363 million users and 5 billion daily clicks of +1 button¹ [1]. OSN users are often not aware of the size and the nature of the audience viewing their profiles, and therefore they may reveal more information than what is appropriate to be viewed publicly. For example, 72% of teens have a social networking profile and nearly half (47%) of them have a public profile accessible by anyone [2]. 15% of Americans have never checked their social networking privacy and security account settings [3]. As a result of lack of privacy awareness, OSNs may often generate a number of privacy related threats for the users. It is reported that the top two social media sites that stalkers use

¹Google+'s +1 button is similar to the like button of Facebook. Users can click the +1 button to show that they like or agree with something, such as photos, videos and news.

are Facebook (16%) and Twitter (3%) [4]. In this dissertation, a quantitative privacy risk analysis framework is proposed. This framework can be used to educate OSN users and raise their privacy protection awareness. It is also able to provide recommendations to improve OSN users' privacy level thus reducing privacy risks.

From the *security* perspective, online product/service review system is one of the most vulnerable systems in social media. First, online product/service has a huge market and it is still growing. More than 80% of the Internet users have used the Internet to make a purchase, and more than 50% of the online shoppers shopped online more than once [5]. E-commerce sales in the US are predicted to grow from \$263 billion in 2013 to \$414 billion in 2018 [6]. Second, online product/service review plays a critical role in online market. There are 74% of consumers that relying on social media to guide their purchases [7]. A one-star rating increase can bring a 5-9% increase of revenue to online sellers [8]. Since there are huge profits of online markets and the purchasing decisions dramatically depend on the product/service reviews, it is highly possible that firms and retailers at the online marketplace may create fake reviews to mislead online customers. In this dissertation, a new angle of fake review detection is introduced. The fake review detection approach based on the new angle can capture dishonest signals that the existing approaches would miss. The success of ERO analysis can improve the trustworthiness of product reviews, can help customers make better decisions, can suppress the fake review market, and therefore can help us build a healthy competitive online market.

1.1 Online Social Network Privacy

1.1.1 Privacy Threats

As the OSNs emerge, people are facing new critical privacy issues. People participate OSN activities by sharing personal data, such as photos, videos, travel

plans, comments, etc. The casual posting of personal information on an OSN often creates a permanent record of the user's personal data, and it creates the possibility for the information to be propagated through social network connections. Some information is expected to be accessible only to certain people, usually intimates or acquaintances. However, ONS is a scale-free network, and digital data is easy to copy and store. Therefore, there are unexpected viewers that may obtain our personal data, and sometimes even private information. This can lead to personal information abuse and affect our daily life.

Real life stories of sensitive information leakage via OSN happen frequently. For example, the UK Ministry of Defence staff have leaked confidential information onto social network sites 6 times in 18 months [9]. The Israeli military cancelled a planned raid on a Palestinian village after one of its soldiers posted details of the operation on Facebook [10]. More employers begin to collect potential employees' information using social network. According to a survey released on the EU Data Protection Day [11], information leakage has put people's careers on risk.

1.1.2 Privacy Protection

ONS privacy issue has been attracting public attention. Hasib has categorized the OSN privacy threats into several types [12], such as digital dossier and search engine indexing that makes sensitive information searchable, image retrieval and interpretation, and profile association. Krishnamurthy et al. studied the problem of personally identity information leakage and how it can be misused by third-parties [13]. Livingstone [14] discussed the risks when young people make friends and share personal information to express themselves online.

Solutions are proposed from different perspectives to protect privacy. On-line social network service providers, including Facebook and Google+, let users manage who can access a certain type of information, and LinkedIn hides sensi-

tive information from users that are not connected to the user. Researchers also studied privacy protection from several perspectives. The first type is proposing new structure of online social network. Felt et al. [15] studied and discussed the privacy concerns of social network APIs (application programming interface) for third party. Baden et al. [16] proposed a new type of online social network using encryption to hide user data and allowing user to define privacy policies. Guha et al. [17] proposed an approach to hide user data by mapping real data to fake data. Another type is developing tools or methods to examine and improve current online social network privacy. Fang et al. [18] developed privacy wizards to give user recommendation for privacy setting. Gundecha et al. [19] proposed an approach to identify a user’s vulnerable friends. In this dissertation, I propose to assist users to improve privacy from another perspective, that is, measuring privacy risk by considering both whether related users are aware of privacy protection and whether they are trusted to protect others’ personal information.

1.1.3 Quantitative Privacy Risk Analysis

According to [20], general security threats can be divided into avoidable category and unavoidable category. OSN privacy issue, in the current design of OSN, is unavoidable, due to the publicity of OSN [21]. When risk is unavoidable, we assume an existing risk and attempt to reduce the likelihood of harmful events.

Under the assumption of unavoidable risk, risk analysis becomes extremely important. With an effective risk analysis method, we are able to 1) design secure privacy management, 2) monitor critical data and protect them effectively, 3) make effective privacy policies, and 4) provide valuable analysis data for future estimation [22].

Risk analysis can be performed either quantitatively or qualitatively. According to the National Institute of Standards and Technology (NIST), risk analysis

is defined as “the process of identifying risk, assessing risk, and taking steps to reduce risk to an acceptable level” [23]. From this guideline, quantitative risk analysis plays a critical role as the reduction of risk need to be measured. **The advantages of quantitative risk analysis include** [24]:

- Many metrics can be used to represent and evaluate the risk parameters. This allows a more detailed analysis of the risky events.
- The risk parameters can be expressed in a numerical way, by which people can understand them and make comparison (e.g. human defined units for data importance, threat impact and reputation loss).
- Sophisticated decision-making techniques can be used as the quantitative assessment provides a credible set of parameters.
- The results of the risk analysis process can be expressed in management’s language. This makes it more efficient to help an entity to achieve risk reducing objectives.

In summary, quantitative risk analysis can assist and strengthen privacy protection from both technical perspective and educational perspective.

However, in the OSN privacy research literatures, quantitatively analyzing privacy risk is still not mature. *First*, privacy risk analysis method should consider the key factors of social network, some which are ignored in the current literature. *Second*, the data availability, OSN structure and privacy policy may change. To our best knowledge, the current methods are not sufficiently flexible with the change of OSN environment. *Finally*, the measurement of privacy level is not clear. Some methods use profile visibility as privacy level [19], and some others use number of friends [25]. Therefore, in **manuscript 1**, I introduce a framework, which is called TAPE (**T**rust-**A**ware **P**rivacy **E**valuation), to quantitatively evaluate the privacy

risk in OSNs. Particularly, the reliability analysis tool is utilized. The factors that impact information propagation are divided into two categories and evaluated by two metrics, including privacy awareness and privacy trust. The proposed TAPE framework is also able to make recommendations to individual OSN user to fulfill their privacy improving needs. the **contributions** of this work include:

1. the TAPE framework, which considers privacy leakage through nodes and links separately and utilizes reliability analysis tools, as well as the definition of privacy risk, in a quantitative way;
2. the privacy awareness and privacy trust metrics;
3. a privacy awareness algorithm, which shows a clear advantage over the known algorithm called IRT [26] in the current literature;
4. the sensitivity analysis metric, from which we propose an efficient unfriending strategy.

Our work can help people understand their privacy situations, raise their privacy awareness and thus reduce privacy risks

1.2 Security of Online Review System

1.2.1 Online Review System

Online product/service review is an option for customers who have experiences of using a product or service. A piece of review usually contains a rating value and some text describing the aspects of the product. Examples of online reviews include Amazon product review, Yelp restaurant review, and TripAdvisor hotel review. Online review system, also referred to as *online reputation system*, allows customers to post reviews for products/services, and integrates these reviews to an overall reputation score. This type of score reveals the average shopping experience, such as product quality and customer service. Online review system helps people

make purchase decisions, and hence greatly affect the profit of online retailers. It is reported that 74% of consumers rely on product/service reviews to guide their purchases [7]. A one-star rating increase can bring a 5-9% increase of revenue to online sellers [8]. People are becoming more and more relying on online reviews when evaluating the quality of products, hotels, restaurants, and vacation packages before placing the order. “If you do build a great experience, customers tell each other about that. Word of mouth is very powerful”, said by Jeff Bezos, CEO of Amazon [27].

1.2.2 Fake Review

Online product reviews are becoming more and more important. However, with the increasing popularity of online review systems, such as Amazon, TripAdvisor and Yelp, and the profit growing of online markets [28], malicious users start to abuse the convenience of publishing online reviews and intentionally post low quality, untrustworthy, or even fraudulent reviews. It is reported that sellers at the online marketplace boost their reputation by trading with collaborators [29], and firms post biased reviews to praise their own products or bad-mouth the competitors’ products [30]. According to Yelp official blog [31, 32], Yelp has been using a review filter to hide certain suspicious reviews since 2005, which means Yelp was aware of review fraudulence at the very early stage. Recently, both Yelp and Amazon announced that they have sued several companies in order to block alleged fake reviews on their website [33, 34].

Fake reviews, also referred to as review manipulation and review spam, can inflate or deflate products’ reputation scores, crash users’ confidence in online reputation systems, and eventually undermine reputation-centric online businesses leading to economic loss. Furthermore, there are some situations, in which the review manipulation is even more damaging. For example, *Black Friday* and **Cyber**

Monday shoppers heavily relies on online reviews, because they have to make rush decisions for the products they are not familiar with in order to take advantage of the ‘unusual’ discount that quickly expires. Another example is online reputation of hotels and restaurants. The consumers, who are misled by manipulated hotel ratings, cannot be easily refunded after they purchase these services.

1.2.3 Online Review System Protection

In the literature, researchers propose methods to protect reputation systems from several angles, such as 1) increasing the cost of acquiring multiple user IDs [35], 2) endogenous discounting of dishonest reviews by analyzing the statistic features of the reviews [36], 3) exogenous discounting of dishonest ratings by introducing reputation evaluation of users [36–38], and 4) studying correlation between users and reviews to detect dishonest reviews [39, 40].

There is a great demand to detect fake reviews thoroughly on the reputation system. There are three directions of fake review detection. In the first direction, fake reviews are detected primarily based on review features, such as the standard word and part-of-speech n-gram features [41], and duplicated or near-duplicated review text [42, 43]. In the second direction, dishonest reviewers are detected based on reviewer features, such as through reviewer graph [44], frequency patterns [45], and user correlations [39]. In the third direction, victim products are detected based on the unusual changes of review statistics, such as the jump/drop of average rating [39] and the arrival rate changes [46].

We argue that victim product detection is critically important and under investigation. First, compared to the huge amount of online reviews, the portion of fake reviews is relatively small [47]. Fake review detections can be more efficient if we can focus on only victim products. Second, some reviews may have limited impact while some other may have major impact on the products’ reputation. Victim

product detection can help us focus on reviews that have major impact. Third, online customers often care about the average rating more than specific reviews. According to the statistics made by Social Barrel, on average one consumer consults 11 online reviews before making a purchasing decision [48]. Knowing whether the product is a victim can help customers make better decisions.

In **manuscript 2**, the Equal Rating Opportunity (ERO) principle is introduced, by assuming that the distributions of certain review features should not be related to the rating value. The ERO analysis is also proposed to detect victim products. The **contributions** of this work include:

1. The ERO principle is introduced, which can reveal the fraudulence signal of fake reviews. Importantly, ERO analysis provides a new angle how we consider product reviews, and it can capture fraudulence signals missed by current work
2. The criterion of ERO feature selection is introduced.
3. The ERO analysis is implemented, and the performance is evaluated by comparing with two other common methods and conducting expert reviews.

1.3 Summary

In summary, comprehensively investigating the multiple privacy and security aspects of social media is of great importance. To address the challenges, this dissertation focuses on two critical aspects, *online social network privacy* and *online review system security*. In this dissertation, I provide reasonable answers to the following questions.

- **Q:** How can we evaluate the behaviors of OSN users of spreading information, based on the current data availability?

A: The information propagation is mainly through social connections and

individual user. The traditional sociology researches can address the evaluation of the impact of social connections, based on the strength of social ties for example. In this dissertation, information propagation through individual user is evaluated by two metrics. One is privacy awareness, which indicates a user's privacy protection awareness, describing whether a user is paying attention to her/his own privacy. The other is privacy trust, which indicates how much a user's friends trust her/him in terms of not gossiping their information to others.

- **Q:** How can we detect the fraudulence signal of fake reviews?

A: Fake review is elusive, and the fake review signal is multidimensional. Fake review detection is nontrivial. The existing approaches can work well for some dimensions. We found a new angle that is not getting sufficient attention, and we introduce the Equal Rating Opportunity principle and ERO analysis to detect fake reviews from this new angle.

List of References

- [1] S. Kemp, "Digital, social & mobile worldwide in 2015," 2015, visited on 2015-09-19. [Online]. Available: <http://wearesocial.net/blog/2015/01/digital-social-mobile-worldwide-2015/>
- [2] National Center for Missing & Exploited Children and John Walsh, "Teen online & wireless safety survey: Cyberbullying, sexting and parental controls," 2009.
- [3] National Cyber Security Alliance, McAfee, and Zogby International, "2011 ncsa / mcafee internet home users survey," 2011.
- [4] T. Riman, "Social media and cyber stalking facts — advice and tips on staying protected," 2013, visited on 2015-10-15. [Online]. Available: <http://www.seomworld.com/2013/09/social-media-and-cyber-stalking-facts.html#.Vlov7iCrTGg>
- [5] "How many people shop online?" 2013, visited on 2015-08-23. [Online]. Available: <http://www.cpcstrategy.com/blog/2013/08/ecommerce-infographic/>

- [6] “U.S. online retail sales will grow 57% by 2018; projected growth,” 2014, visited on 2015-03-25. [Online]. Available: <https://www.internetretailer.com/2014/05/12/us-online-retail-sales-will-grow-57-2018>
- [7] S. Bennett, “The rise of social media in e-commerce,” 2014, visited on 2015-11-18. [Online]. Available: <http://www.adweek.com/socialtimes/social-e-commerce/496484>
- [8] Q. Fottrell, “Yelp deems 20% of user reviews ‘suspicious’,” 2013, visited on 2015-08-09. [Online]. Available: http://www.marketwatch.com/story/20-of-yelp-reviews-are-fake-2013-09-24?link=MW_story_latest_news
- [9] S. News. “Mod secrets leaked onto the internet.” 2010. [Online]. Available: <http://news.sky.com/story/753966/mod-secrets-leaked-onto-the-internet>
- [10] B. News. “Israeli military ‘unfriends’ soldier after facebook leak.” 2010. [Online]. Available: <http://news.bbc.co.uk/2/hi/middle-east/8549099.stm>
- [11] EurActiv. “Social networks put careers at risk, survey finds.” 2010. [Online]. Available: <http://www.euractiv.com/Social-networks-careers-risk>
- [12] A. Al Hasib, “Threats of online social networks,” *IJCSNS International Journal of Computer Science and Network Security*, vol. 9, no. 11, pp. 288–93, 2009.
- [13] B. Krishnamurthy and C. E. Wills, “On the leakage of personally identifiable information via online social networks,” in *Proceedings of the 2nd ACM workshop on Online social networks*, Barcelona, Spain, 2009.
- [14] S. Livingstone, “Taking risky opportunities in youthful content creation: teenagers’ use of social networking sites for intimacy, privacy and self-expression,” *New media & society*, vol. 10, no. 3, pp. 393–411, 2008.
- [15] A. Felt and D. Evans, “Privacy protection for social networking APIs,” in *Web 2.0 Security and Privacy*, Oakland, CA, USA, 2008.
- [16] R. Baden, A. Bender, N. Spring, B. Bhattacharjee, and D. Starin, “Persona: an online social network with user-defined privacy,” in *Proceedings of the ACM SIGCOMM 2009 conference on Data communication*, Barcelona, Spain, 2009.
- [17] S. Guha, K. Tang, and P. Francis, “NOYB: privacy in online social networks,” in *Proceedings of the first workshop on Online social networks*, Seattle, WA, USA, 2008.
- [18] L. Fang and K. LeFevre, “Privacy wizards for social networking sites,” in *Proceedings of the 19th international conference on World Wide Web*, Raleigh, NC, USA, 2010.

- [19] P. Gundecha, G. Barbier, and H. Liu, “Exploiting vulnerability to secure user privacy on a social networking site,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, San Diego, CA, USA, 2011.
- [20] S. E. Schechter, “Computer security strength and risk: A quantitative approach,” Ph.D. dissertation, Harvard University, Cambridge, MA, USA, 2004.
- [21] S. Strauß and M. Nentwich, “Social network sites, privacy and the blurring boundary between public and private spaces,” *Science and Public Policy*, vol. 40, no. 6, pp. 724–732, 2013.
- [22] H. In, Y.-G. Kim, T. Lee, C.-J. Moon, Y. Jung, and I. Kim, “A security risk analysis model for information systems,” in *Systems Modeling and Simulation: Theory and Applications*. Springer Berlin Heidelberg, 2005, vol. 3398, pp. 505–513.
- [23] G. Stoneburner, A. Goguen, and A. Feringa, “Risk management guide for information technology systems,” *Nist special publication*, vol. 800, no. 30, 2002.
- [24] M. Hamdi and N. Boudriga, “Computer and network security risk management: theory, challenges, and countermeasures,” *International Journal of Communication Systems*, vol. 18, no. 8, pp. 763–793, 2005.
- [25] R. Abdulrahman, S. Alim, D. Neagu, and M. Ridley, “Algorithms for data retrieval from online social network graphs,” in *IEEE 10th International Conference on Computer and Information Technology*, Bradford, UK, 2010.
- [26] S. Guo and K. Chen, “Mining privacy settings to find optimal privacy-utility tradeoffs for social network services,” in *International Conference on Privacy, Security, Risk and Trust and International Confernece on Social Computing*, 2012.
- [27] “Online extra: Jeff bezos on word-of-mouth power,” 2004, visited on 2014-09-19. [Online]. Available: <http://www.bloomberg.com/bw/stories/2004-08-01/online-extra-jeff-bezos-on-word-of-mouth-power>
- [28] “Final pre-christmas push propels u.s. online holiday season spending through december 26 to record \$30.8 billion,” 2010, visited on 2015-01-20. [Online]. Available: <http://ir.comscore.com/releasedetail.cfm?ReleaseID=539354>
- [29] J. Morgan and J. Brown, “Reputation in online auctions: The market for trust,” *California Management Review*, vol. 49, no. 1, pp. 61–81, 2006.
- [30] A. HARMON, “Amazon glitch unmasks war of reviewers,” 2004, visited on 2015-01-20. [Online]. Available: <http://www.nytimes.com/2004/02/14/us/amazon-glitch-unmasks-war-of-reviewers.html>

- [31] Yelp, “Yelp’s review filter explained,” 2010, visited on 2015-03-21. [Online]. Available: <http://officialblog.yelp.com/2010/03/yelp-review-filter-explained.html>
- [32] Yelp. “Why yelp has a review filter.” Visited on 2015-03-21. 2009. [Online]. Available: <http://officialblog.yelp.com/2009/10/why-yelp-has-a-review-filter.html>
- [33] J. STEMPEL. “Amazon sues to block alleged fake reviews on its website.” Visited on 2015-06-02. 2015. [Online]. Available: <http://www.reuters.com/article/2015/04/10/us-amazon-com-lawsuit-fake-reviews-idUSKBN0N02LP20150410>
- [34] A. PICCHI. “Yelp sues companies promising positive reviews.” Visited on 2015-08-19. 2015. [Online]. Available: <http://www.cbsnews.com/news/yelp-sues-companies-promising-positive-reviews/>
- [35] M. Abadi, M. Burrows, B. Lampson, and G. Plotkin, “A calculus for access control in distributed systems,” *ACM Trans. Program. Lang. Syst.*, vol. 15, no. 4, pp. 706–734, 1993.
- [36] A. Josang, R. Ismail, and C. Boyd, “A survey of trust and reputation systems for online service provision,” *Decision Support Systems*, vol. 43, no. 2, pp. 618 – 644, 2007.
- [37] P. Laureti, L. Moret, Y.-C. Zhang, and Y.-K. Yu, “Information filtering via iterative refinement,” *EPL (Europhysics Letters)*, vol. 75, no. 6, p. 1006, 2006.
- [38] B. Yu and M. P. Singh, “An evidential model of distributed reputation management,” in *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1*, ser. AAMAS ’02. ACM, 2002, pp. 294–301.
- [39] Y. Liu and Y. Sun, “Anomaly detection in feedback-based reputation systems through temporal and correlation analysis,” in *2010 IEEE Second International Conference on Social Computing (SocialCom)*, Aug 2010, pp. 65–72.
- [40] C. Dellarocas, “Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior,” in *Proceedings of the 2Nd ACM Conference on Electronic Commerce*, ser. EC ’00. ACM, 2000, pp. 150–157.
- [41] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, “Finding deceptive opinion spam by any stretch of the imagination,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT ’11, 2011, pp. 309–319.

- [42] R. Y. K. Lau, S. Y. Liao, R. C.-W. Kwok, K. Xu, Y. Xia, and Y. Li, “Text mining and probabilistic language modeling for online review spam detection,” *ACM Trans. Manage. Inf. Syst.*, vol. 2, no. 4, pp. 25:1–25:30, 2012.
- [43] E. Choo, T. Yu, M. Chi, and Y. Sun, “Revealing and incorporating implicit communities to improve recommender systems,” in *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, ser. EC ’14. ACM, 2014, pp. 489–506.
- [44] G. Wang, S. Xie, B. Liu, and P. Yu, “Review graph based online store review spammer detection,” in *2011 IEEE 11th International Conference on Data Mining (ICDM)*, Dec 2011, pp. 1242–1247.
- [45] A. Mukherjee, B. Liu, J. Wang, N. Glance, and N. Jindal, “Detecting group review spam,” in *Proceedings of the 20th International Conference Companion on World Wide Web*, ser. WWW ’11. ACM, 2011, pp. 93–94.
- [46] Y. Yang, Y. Sun, S. Kay, and Q. Yang, “Securing rating aggregation systems using statistical detectors and trust,” *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 4, pp. 883–898, 2009.
- [47] C. Lai, K. Xu, R. Lau, Y. Li, and L. Jing, “Toward a language modeling approach for consumer review spam detection,” in *2010 IEEE 7th International Conference on e-Business Engineering (ICEBE)*, 2010, pp. 1–8.
- [48] S. H. Dolor, “30 digital marketing statistics you shouldnt miss,” 2013, visited on 2015-11-18. [Online]. Available: <http://socialbarrel.com/30-digital-marketing-statistics-shouldnt-miss-infographic/53453/>

CHAPTER 2

“Online Social Networks Privacy Study Through TAPE Framework”

by

Yongbo Zeng¹, Yan (Lindsay) Sun², Liudong Xing³, and Vinod Vokkarane⁴

is published in IEEE Journal of Selected Topics in Signal Processing, 2015

¹PhD Candidate, Department of Electrical, Computer & Biomedical Engineering, The University of Rhode Island, Kingston, RI 02881. Email: yongbozeng@ele.uri.edu

²Professor, Department of Electrical, Computer & Biomedical Engineering, The University of Rhode Island, Kingston, RI 02881. Email: yansun@ele.uri.edu

³Professor, Department of Electrical & Computer Engineering, The University of Massachusetts Dartmouth, MA. Email: lxing@umassd.edu

⁴Associate Professor, Department of Electrical and Computer Engineering, The University of Massachusetts Lowell, MA, Email: vinod.vokkarane@uml.edu

2.1 Abstract

While personal information privacy is threatened by online social networks, researchers are seeking for privacy protection tools and methods to assist online social network users. In this manuscript, we propose a Trust-Aware Privacy Evaluation framework, called TAPE, aiming to address this problem. Under the TAPE framework we investigate how to quantitatively evaluate the privacy risk, as a function of people’s awarenesses of privacy risks as well as whether people can be trusted by their friends to protect others’ personal information. Simulations are performed to illustrate the key concepts and calculations, as well as the advantages of TAPE. Based on the TAPE framework, we also propose an unfriending strategy in terms of privacy protection, which outperforms other existing unfriending strategies.

2.2 Introduction

With the emergence of **Online Social Networks (OSN)**, people are facing critical privacy risks. In OSN, personal information can be abused, which will put users into risks. Researchers identified OSN privacy issues as two categories, inadvertent disclosure of personal information, and stalking or backtracking [1, 2]. Krishnamurthy et al. studied the problem of personal identity information leakage and how it can be misused by third parties [3]. This kind of information is able to distinguish an individual’s identity either alone or when being combined with other information that is linked to a specific individual, and its leakage will lead to identity theft. Livingstone [4] demonstrated the risks when young people make friends and share personal information to express themselves online. Real life stories of sensitive information leakage in OSNs happen frequently. For example, most employers began to collect potential employees’ information using social networks. According to a survey released on the EU Data Protection Day [5], privacy leakage

had put people’s careers on risk.

In the current commercial OSN design, privacy risk is unavoidable, due to the publicity of OSN [6, 7]. In order to benefit from the convenience of OSNs, people share personal information with friends, which makes privacy leakage possible. When privacy risk is unavoidable, we assume a risk and attempt to reduce the likelihood of harmful events. Under the assumption of unavoidable risk, risk analysis becomes extremely important. According to the National Institute of Standards and Technology (NIST), risk analysis is defined as “the process of identifying risk, assessing risk, and taking steps to reduce risk to an acceptable level” [8]. In [9], In et al. discussed the advantages of risk analysis, including designing secure privacy management, monitoring and protecting critical data, making privacy policies, etc.

Risk analysis can be performed either quantitatively or qualitatively. Quantitative risk analysis plays a critical role. The advantages of quantitative risk analysis are discussed in [10], such as using metrics to evaluate risk parameters, analyzing risk events, making sophisticated decisions.

Solutions for addressing privacy issue include educational aspect and technical aspect. For example, in [11], Gundecha et al. studied privacy issues and protection recommendations, for the purpose of educating OSN users and raising their privacy awareness. The technical aspect includes managing privacy setting [12] and adopting new architectures to build OSNs [13] [14]. There are also some other categorizations in the literature. For example, in [15], Jeckmans et al. categorized privacy research into 5 categories, including raising user awareness, law and regulations, personal information anonymization, perturbing user information, and data encryption. We attempt to solve privacy issue from another perspective – providing a quantitative privacy risk analysis framework for OSN users and researchers. In the OSN privacy research literatures, quantitatively analyzing privacy risk is

still not mature, as we discuss in Section 2.3. Therefore, we propose a framework in this manuscript to quantitatively evaluate the privacy level of OSN users. We believe that the proposed framework can help people to understand their privacy situations, raise their privacy awareness and thus reduce privacy risks.

Quantitatively evaluating privacy level in OSN is a challenging task. First, quantitative user privacy level is not a well defined concept in OSNs. Second, human users play an important role in the personal information leakage. It is complicated to predict an individual user’s behavior. Third, personal information can be propagated through both online and offline media by many ways, such as chatting, emails, instant messages, Facebook postings, picture postings, tweets on Twitter, etc. Fourth, it is extremely difficult, if not impossible, to obtain the ground truth about a user’s privacy level, with which the quantitative evaluation results can be compared.

In this manuscript, we address the *first challenge* by proposing quantitative definition of privacy risk, based on privacy hazard and its probabilities. This quantitative measurement will lead to the privacy level calculation tools, which were originally proposed in the reliability analysis field. To address the *second and third challenge*, we have to consider the availability of social data. Since nobody can monitor the users’ all communication behaviors (online and offline), researchers have to work on limited data, which can be obtained with reasonable costs. In this work, Facebook privacy setting is used as the primary data source. We also focus on the ‘word-of-mouth’, which is the primary drive of OSN information diffusion [16]. Although other privacy leakage scenarios, which we discuss in Section 2.7.7, are not considered in this work, the proposed concepts, including privacy awareness and privacy trust, can be extended to those scenarios. For the *fourth challenge*, due to the lack of ground truth of users’ privacy level, we compare the proposed

scheme with some existing approaches, such as privacy concern model in [17] and vulnerability analysis in [11]. In addition, Monte Carlo simulation is employed to verify the results of privacy risk evaluation.

In this manuscript, we propose a TAPE (Trust-Aware Privacy Evaluation) framework for quantitatively evaluating users' privacy level in OSNs. The TAPE framework contains several novel aspects.

- It finds the similarity between the reliability analysis in wireless sensor networks (WSN) and the privacy risk estimation in OSNs. It sets up the stage for utilizing reliability analysis tools for privacy analysis.
- It considers the privacy leakage through nodes (i.e. users) and through links (i.e. friend connections) separately. Here, the privacy leakage through nodes mainly depends on the users' behavior, and we define two metrics in TAPE to estimate it. The first one reflects whether one respects others' privacy, and it is named as *Privacy Awareness*. The other one reflects how much one's friends trust her/him in terms of not gossiping their information to others, and it is named as *Privacy Trust*. The privacy leakage through a link mainly depends on the relationship between the two users, in terms of whether one paying attention to the other' personal information.
- It proposes the desirable properties of privacy awareness and privacy trust metrics, as well as specific ways to calculate them under the guidance of trust management theory. It is the first time that the privacy trust concept has been used in evaluating privacy level in OSN.

Besides privacy risk estimation, the TAPE framework has the ability to conduct **sensitivity analysis** for friend links, which is similar to the concept of vulnerability in [11]. Through the sensitivity analysis, an OSN user can understand

how much his/her privacy level is affected by a particular friend connection. The sensitivity analysis yields a practical way to improve OSN users' privacy level.

As a summary, the **contributions** of this work include:

1. the TAPE framework, which considers privacy leakage through nodes and links separately and utilizes traditional reliability analysis tools, as well as the definition of privacy risk, in a quantitative way;
2. the privacy awareness and privacy trust metrics;
3. a privacy awareness algorithm, which shows a clear advantage over the know algorithm called IRT [17] in the current literature;
4. the sensitivity analysis metric, from which we propose an efficient unfriending strategy.

This manuscript is organized as follows. Related work is discussed in Section 2.3. TAPE framework is described in Section 2.4, followed by discussion of information spreading probability algorithms and the proposed algorithms in Section 2.5. Privacy assessment and sensitivity analysis metric are presented in Section 2.6. Experiment results and conclusion are presented in Section 2.7 and Section 2.8 respectively.

2.3 Related Work

Privacy in OSN have attracted many attentions. OSN service providers allow users to manage who can access which information (e.g. in Facebook and Google+), and to hide sensitive information to non-connected users (e.g. in LinkedIn). Researchers studied privacy protection from two directions. Along the *first direction*, fundamental changes to the current design of OSN were suggested to enhance users' privacy. Felt et al. [14] studied and discussed the privacy concerns of social

network APIs for third parties. Guha et al. [18] proposed an approach to hide user data by mapping real data to fake data. Within the first direction, “Privacy by Design” (PbD) is an important approach. In [19], Wolf et al. operationalized the concept of PbD through the process of design and development of OSN, and several social requirements of OSN were identified to optimize the privacy from a user perspective. Encryptions are usually used when adopting PbD. For example, in [13], Baden et al. proposed a new type of OSNs by using attribute-based encryption to hide user data, in which symmetric keys are used to encrypt messages and only the designated friend groups can decrypt the messages. In [20], Erkin et al. proposed to use homomorphic encryption and multi-party computation techniques to hide privacy-sensitive data from the service provider in a recommender system, without losing the significant usability of data. The *second direction* is developing privacy protection tools based on existing OSNs. For example, Fang et al. [12] developed privacy wizards to give user recommendation for privacy setting, and Gundecha et al. [11] proposed an approach to identify a user’s vulnerable friends. In this manuscript, we propose to assist users’ privacy protection by providing quantitative evaluation of privacy risk and conducting sensitivity analysis for friend links. Our work belongs to the second direction.

There have been several quantification models for privacy evaluation in OSN. Alim et al. [21] proposed a vulnerability quantification model which consists of three components: individual, relative and absolute vulnerabilities. They examined the visibility of OSN users’ profiles and computed the clustering coefficient to compose individual vulnerability. Based on individual vulnerability, relative vulnerability and absolute vulnerability were calculated. Besides privacy risk evaluation, friend vulnerability analysis, also referred to as sensitivity analysis in this manuscript, is considered to be a good way to improve personal privacy. Abdulrah-

man et al. proposed a node vulnerability metric [22] and a multi-agent vulnerability analysis [23] based on the friendship graph of MySpace. Vulnerability index (V-Index) was proposed to measure how vulnerable an OSN user is based on her/his friends' privacy setting [11]. Privacy setting and its implications were considered as a primary factor in the existing models. In this manuscript, we consider privacy setting as one of the primary factors. The implications of privacy setting are represented as two metrics – privacy awareness and privacy trust. Besides privacy setting, the TAPE framework is able to adopt social tie analysis approaches when implementing the module of link information spreading probability. The network topology and information diffusion patterns are also considered.

The proposed work is also related to information diffusion in OSN. Gruhl et al. [24] studied the dynamics of information spreading in weblogs. Adar et al. [25] demonstrated a technique for inferring information propagation through a blog network by applying epidemic models of information spreading. Cha et al. [16] studied social cascades over Flickr social network. Researchers also attempt to build mathematical model to solve problems of information diffusion in online social network, such as [26–28]. In addition, there are literatures discussing the relationship between tie strength and information propagation [29], which is related to the information spreading probability that is discussed in this manuscript. Different from the previous information diffusion work, the proposed TAPE framework considers information diffusion in the context of privacy protection, which requires different set of features and considerations.

2.4 Trust-aware Privacy Evaluation Framework

In this section, the TAPE framework is discussed in details. We first define privacy risk from the perspective of information diffusion. The binary decision diagram (BDD) which was commonly used for system reliability analysis is employed

to calculate privacy risks. The concepts of node information spreading probability and link information spreading probability are proposed.

2.4.1 Acronyms

OSN Online Social Network

PIO Personal Information Owner

UD Undesirable Destination

UG Undesirable Group

PA Privacy Awareness

IPA Individual Information Privacy Awareness

PT Privacy Trust

ISP Information Spreading Probability

NISP Node Information Spreading Probability

LISP Link Information Spreading Probability

WSN Wireless Sensor Network

BDD Binary Decision Diagram

BM Birnbaum's Measure

IRT Item Response Theory

2.4.2 Notations

I_j Alice's type j personal information

UG Alice's undesirable group related to I_j

UD_i Alice's i th undesirable destination in UG

V_j Privacy risk for I_j

Z_j Privacy leakage hazard for I_j

L_j Privacy leakage probability for I_j

$ISP(.)$ Calculation of ISP

S The set of Alice's privacy setting

s_j Alice's privacy setting for I_j

PA_u User u 's PA

$G_{PA}(.)$ Calculation of PA

$rank_{u,j}^+$ Proportion of users whose privacy setting for I_j is looser than user u

$rank_{u,j}^-$ Proportion of users whose privacy setting for I_j is tighter than user u

$\langle A, B \rangle$ Friend link between A and B

PT_u User u 's PT

PT_{u,f_i} PT evaluation based on friend f_i 's recommendation

$G_{PT}(.)$ Calculation of PT

$T_{A,B}$ Evaluation of how much A trusts B

R_u^+ Positive recommendations for user u

Figure 1: Online social network of Example 1.

2.4.3 Online Social Network Privacy

Some OSNs (e.g. Facebook and LinkedIn) encourage people to use real names and upload personal information onto a page known as ‘Profile’. Such personal information is often accessible by friends directly, and can even flow to thousands of other people through retweet (e.g. on Twitter), sharing (e.g. on Facebook) and online communication (e.g. chatting). The privacy concern in OSNs is well known, but how can we define the privacy risk in a quantitative way?

Before discussing quantification of privacy risk, we first look at two examples.

Example 1. *Alice is a student, and she posted a piece of comment complaining her teacher Cris on her social network site. Alice does not want Cris and other teachers in the same department to know the comment. Fig 1 shows the example social network.*

Example 2. *Alice posted a photo, and she does not want anyone, except her friends, to see this photo.*

Generally, in some scenarios, we want some personal information to be known only by friends, and in some other scenarios we don't want certain personal information to be viewed by specific people [2]. In Example 1, the personal information concerned by Alice is her comment on Cris, and in Example 2, the personal information is her photo. It is clear that an user has different types of personal information, and that the privacy concerns depend on the particular type of personal information. We introduce the notation I_j^u to denote user u 's type j personal information. Without loss of generality, we present the framework in the context of protecting Alice's privacy, i.e. $u = \text{"Alice"}$. Alice is also referred to as the **personal information owner (PIO)**. In the rest of the manuscript, for simplicity, we use I_j to represent I_j^{Alice} .

It is noted that privacy concerns are related to the "undesirable viewers". We define the concept of **Undesirable Group (UG)** of I_j , denoted by UG_j , as follows. If Alice does not want her information I_j to be seen by user u_i , then u_i is put into UG_j , where u_i is also called **Undesirable Destination (UD)** of I_j . In Example 1, Alice's UG is {Cris}. In Example 2, Alice's UG contains all users except her friends.

In other words, if I_j flows to any UD, Alice considers her privacy of I_j being violated and **privacy leakage** happens. In the rest of the manuscript, for simplicity, we use UG to represent UG_j^{Alice} .

Information leaking to different persons has different potential risks to the PIO. Without defining UD, this difference cannot be captured. The privacy definition based on UD is a more generalized definition. In this definition, users are classified into 4 types:

1. personal information owner (e.g. Alice),
2. users who are allowed to access personal information according to the privacy

setting (e.g. Alice’s friends),

3. users to whom the exposure of personal information causes damage (i.e. the undesirable group)
4. users not belonging to the above three types.

In the existing work, people usually assume that there are no type 4 users, such as in [11]. The definition of privacy leakage in this manuscript becomes the traditional definition as long as the UD is defined as the complement set of the type 1 and type 2 users. Our definition can also handle the cases that Alice only concerns that the privacy leaks to a specific set of users, as seen in Example 1. In other words, our definition can capture the fact that privacy leaking to different persons has different damage to the PIO. Such difference is usually not captured by the privacy setting alone.

2.4.4 Privacy Risk and Related Concepts

With the proposed TAPE framework, we aim to answer two questions: 1) Can we measure the probability of personal information leakage as a measurement of privacy risk in OSN? 2) How is the personal information leakage related to privacy risk? In this subsection, we first introduce the key concepts of the TAPE framework.

In [15], privacy is considered as keeping a piece of information in its intended scope. In TAPE, the leakage of personal information I_j occurs when any users in the undesirable group UG_j view I_j . Here the undesirable group is the same as the unintended scope in [15]. We assume that I_j can only be obtained through online information diffusion, which only occurs through friend connections. This assumption is a result of the limitation of data, as discussed in Section 3.2. In the future, if more data are available, such as cell phone contact data, this assumption

Figure 2: Core structure of the TAPE framework

can be revised. Due to this assumption, the UG in Example 2 can be simplified as $\{all\ of\ Alice's\ 2-hop\ neighbors\}$. We define **privacy leakage probability** of I_j , denoted by L_j , as the probability that at least one UD views I_j through information diffusion in the OSN.

$$L_j = Pr\{privacy\ leakage\ happens\} \quad (1)$$

In statistics, the notion of risk is often modeled as the expected value of an undesired outcome [30]. That is

$$\begin{aligned} Risk = & (probability\ of\ the\ accident\ occurring) \\ & \times (expected\ loss\ in\ case\ of\ the\ accident) \end{aligned} \quad (2)$$

In the context of OSN, we argue that privacy risk of information I_j , denoted by V_j can be computed as

$$V_j = L_j \cdot Z_j \quad (3)$$

where L_j is privacy leakage probability as defined in Equation 1 and Z_j describes the expected loss/damage in case of privacy leakage. In this manuscript, we also use another term “privacy level” to describe an individual’s privacy, and obviously, the lower privacy risk is, the higher the privacy level is. In TAPE, Z_j is referred to as **privacy leakage hazard** and is normalized within interval $[0, 1]$. We argue that Z_j should be determined by the PIO (e.g. Alice) when the damage of the privacy leakage is subjective. For instance, in Example 1, Alice may be the best person who determines the damage if Cris saw her complain on Facebook? In many cases, PIO is often the best person to estimate the damage/loss of privacy leakage. In some other cases, PIO may not have the knowledge to determine the damage. For example, in Example 2, Alice may not understand the consequence of revealing her photo to strangers. In such case, Alice needs assistance to determine the damage, such as the privacy leakage problem study in [3]. In this manuscript, we simply assume that Z_j can be provided by PIO. In the rest of this manuscript, when we

compare privacy risks, Z_j is considered as constant 1. Based on this assumption, the privacy leakage probability L_j is equivalent to privacy risk V_j . *The core task in TAPE is to estimate the privacy leakage probability L_j .*

2.4.5 Toward Privacy Leakage Probability Estimation

In TAPE, a social network is represented by an undirected diagram. OSN users are the nodes, denoted as u_i , and their friend connections are the links, denoted as $\langle u_m, u_n \rangle$. As discussed in the previous section, personal information can be diffused to unintended recipients through friend links. It is important to point out that the existence of a link does not necessarily mean the personal information will be transmitted through it. For instance, in Example 1, Alice posted a piece of comment, but some of her friends may not read it. Here are three typical situations.

- Alice’s friend Bob does not pay attention to Alice’s posting at all. Alice’s comment does not disseminate to Bob through the link between Alice and Bob.
- Alice’s friend Bob pay attention to Alice’s posting and read it. Alice’s comment disseminates to Bob through the link between Alice and Bob. Then, Bob respects Alice’s privacy and does not tell others about Alice’s comment. In this case, Alice’s comment does not disseminate to others through Bob.
- Bob reads Alice’s comment, and retweets it. Such retweeting can be seen by all Bob’s friends. In this case, Alice’s comment disseminates to Bob, and it is possible to disseminate to others through Bob.

We argue that the privacy leakage probability estimation problem can be decomposed into two tasks.

Table 1: Concepts mapping

Reliability Analysis for WSN	TAPE
System reliability	Privacy leakage probability
Reliability graph	Social graph
Source node	Personal information owner (PIO)
Destination	Undesirable destination (UD)
Node/edge failure probability	1 - node/link information spreading probability (ISP)

1. **Task 1:** The first task is to estimate the probability whether a user’s personal information will be disseminated through a particular component c (a link or a node). In this work, such probabilities are referred to as **information spreading probabilities** (ISP), denoted by $ISP(c)$. The ISP of the link between Alice and Bob depends on factors such as whether Alice and Bob are good friends, whether Bob actively communicates with Alice in OSN, and whether the information is interesting enough to catch Bob’s attention. The ISP of a node is determined by complicated factors, ranging from knowledge to personality, which is extremely difficult to quantify or even understand.
2. **Task 2:** The second task is to compute the probability of privacy leakage (i.e. L_j), given the network topology, the information spreading probabilities of links and nodes, the PIO (i.e. Alice), and the UG.

In the rest of this section, we first discuss the solution to the second task (Section 2.4.6), and then present the solution to the first task (Section 2.5). Fig 2 shows the core structure of the TAPE framework.

2.4.6 Privacy Analysis and Reliability Analysis

When investigating information diffusion in OSN, we found reliability graph, which has been used as one of the reliability analysis tools (e.g. WSN reliability),

(a) Wireless sensor network example

(b) Online social network example

Figure 3: Similarity between WSN and OSN. In 3a, Sensor A detects fire, and the detection will be sent to the server through other sensor nodes. In 3b, Alice (PIO) feels her photos are improper to be viewed by Eve (UD).

can be adapted to solve the problem.

In a reliability analysis problem, the system is represented by a **reliability graph**, whose links and nodes are assigned **failure probabilities**. The system has a **source** node and a **sink** node that is usually a station. If there is no path from the source to the sink can be established, the system fails. For example, in a WSN, the nodes are sensors, and the links are the communication channels. A sensor's failure probability depends on its battery, environment temperature, work

Figure 4: An OSN information diffusion example.

load, etc. A communication channel’s failure probability depends on distance, environment noise, etc. In the context of reliability analysis, one often needs to estimate the probability that there is at least one path can be established from the source sensor to the destination sensor [31].

In the TAPE framework, we have defined the information spreading probability for nodes and links in the previous section. This concept is kind of “opposite” to the failure probability. For example, if node A fails to forward data to its neighboring nodes with probability x , node A’s failure probability is x in the context of WSN reliability analysis, whereas this node’s information spreading probability is $1 - x$ in the context of privacy analysis. The goal of WSN is to transmit data successfully, whereas the goal of privacy protection is to prevent personal information from propagation. Therefore, in the TAPE framework, we can also define failure probability of nodes/links as $1 - ISP$. We propose to use the binary decision diagram (BDD) method, which is commonly used in reliability analysis [31–33], to solve Task 2 described in Section 2.4.5. Table 1 shows the important concepts in TAPE, as well as the concepts mapping.

A BDD is a directed acyclic graph created based on Shannon’s decomposition. It is an efficient tool to manipulate boolean expressions. For example, in Fig 4, Alice is PIO and Bob is UD. All nodes and links are assigned ISPs. In order to calculate the information leakage probability L_j , we first use a boolean expression to represent L_j .

Figure 5: BDD graph of the example in Fig 4

$$\begin{aligned}
 L_j = & ISP(e_1)ISP(n_1)(ISP(e_4) + ISP(e_3)ISP(n_2)ISP(e_5)) \\
 & + ISP(e_2)ISP(n_2)(ISP(e_5) + ISP(e_3)ISP(n_1)ISP(e_4))
 \end{aligned} \tag{4}$$

Then, a BDD graph is constructed based on the reliability expression. The BDD graph is a binary tree (Fig 5), each sub-tree is considered to be a sub-expression. The left sub-tree of a BDD node represents the expression when the node successfully spreads information. The right sub-tree represents the expression when the node fails to spread information. When traversing from the root to a leaf node, if the leaf node is a left child, then it gives a information leakage case. Based on the BDD diagram, we can evaluate L_j using a recursive method. The details of the BDD approach can be found in [31].

When BDD method is utilized in the OSN privacy problem, one of the most challenging issue is the computational cost. It is noticed that the size of the BDD graph increase exponentially as the size of the network. The size of OSNs are too large to make an efficient BDD calculation. In this work, BDD is employed to compute the probability of information diffusion after modification. Due to the

large size of the social network and the high computation cost of BDD, we adopt a **reduced BDD algorithm**. In particular, we set the maximum traversing depth as k times the number of hops between PIO and UD. For example, when $k = 2$ and the UD is 3 hops away from the PIO, the branches longer than 6 (3×2) are discarded from the BDD graph. In Section 2.7, we set $k = 2$.

2.4.7 Summary

By studying the similarity between the reliability analysis in WSN and the privacy risk estimation, we modify the BDD method to evaluate information leakage probability. The concept of node ISP and link ISP are developed. The core structure of the TAPE is shown in Fig 2. As a summary, TAPE is presented as a framework to solve task 2 described in Section 2.4.5. In Section 2.5, we discuss details of ISP calculation. Particularly, the metrics of privacy awareness and privacy trust are proposed for node ISP calculation.

2.5 Information Spreading Probability Algorithms

While most social network information diffusion models consider the impact of nodes and links together [25], we argue that information propagation through nodes and through links should be considered separately. This is why we define **information spreading probability of node (NISP)**, also referred to as node ISP, and **information spreading probability of link (LISP)**, also referred to as link ISP, which can better describe the information diffusion process. NISP is the probability that a node will spread others' information, and LISP is the probability that a link will be in the path of information diffusion. NISP and LISP imitate the nature human communication process in the real world (i.e. offline social network).

- NISP describes the probability of speaking, i.e. talking about others.

- LISP describes the probability of listening, i.e. hearing what is said.

In this section, we focus on the algorithm of NISP, followed by a brief introduction of the LISP algorithms proposed in literatures.

2.5.1 Node Information Spreading Probability (NISP)

Evaluating NISP of a person is very challenging, because it is related to one's knowledge and personality. In the offline social network, we probably can estimate the NISP of a person based on experiences if we know this person well. Obviously, such estimation can be biased and limited, and most importantly cannot be applied in OSNs due to data limitation. Instead of resolving a challenging problem in social science, we propose to examine NISP based on the quantitative data available in OSNs.

In particular, we propose two metrics that should be used to estimate NISP – privacy awareness and privacy trust.

Privacy Awareness

The first metric is **privacy awareness (PA)**, which depends on a user's privacy setting. We argue that privacy setting reflects a user's privacy protection awareness, describing whether a user is paying attention to his/her own privacy. There are many different ways to compute a user's PA. In TAPE, *PA evaluation is a module*. The input is a set of the user's privacy setting, which is represented as $S_u = \{s_{u,j} | j = 1, 2, \dots, J\}$ where u is the user and $s_{u,j}$ is the privacy setting for information I_j . Privacy setting has options $\{v_1, v_2, \dots, v_M\}$, in which v_n is a looser setting than v_m for $n < m$, and correspondingly, $s_{u,j}$ can have a value in $\{1, 2, \dots, M\}$. For example, in Facebook, $v_1 = \text{'everyone'}$, $v_2 = \text{'networks and groups'}$, $v_3 = \text{'friends of friends'}$, $v_4 = \text{'friends'}$, and $v_5 = \text{'self'}$. Let I_1 be the photo in Example 2. Alice's privacy setting for I_1 is 'friends', i.e. $s_{Alice,1} = 4$. Without

Table 2: Desirable properties of PA algorithms

Special Cases	$rank_1^+$	$rank_1^-$	Desirable PA value when normalizing to $[0, 1]$
Case 1: Alice’s privacy setting is looser than all others’	0	≈ 0	$PA_{Alice} \approx 0$
Case 2: Alice’s privacy setting is tighter than all others’	≈ 1	0	$PA_{Alice} \approx 1$
Case 3: Everyone has the same privacy setting	0	0	$PA_{Alice} = 0.5$
Case 4: Many users (including Alice) have loose setting, and a few users have tight setting	0	small	$PA_{Alice} < 0.5$, and it should be small, but not too small because most people share the same opinion as Alice.
Case 5: A few users (including Alice) have tight setting, and many users have loose setting	big	0	$PA_{Alice} > 0.5$, and it should be higher than PA_{Alice} in case 7.
Case 6: A few users (including Alice) have loose setting, and many users have tight setting	0	big	PA_{Alice} should be smaller than PA_{Alice} in case 4.
Case 7: Many users (including Alice) have tight setting, and a few users have loose setting	small	0	PA_{Alice} should be high, but not too high because most people share the same opinion as Alice.

loss of generality, in the rest of this manuscript, we use S and s_j to represent Alice’s privacy setting set and privacy setting respectively.

Let PA_u represent PA of user u and G_{PA} represent the adopted PA calculation, then

$$PA_u = G_{PA}(S_u) \quad (5)$$

The TAPE framework can accommodate many PA algorithms. However, what are the design criteria for PA algorithms? Based on the possible distributions of privacy setting and Alice’s possible adoptions, we identified seven special cases and the desirable PA values in these special cases in Table 2, which serves as a guidance for the PA algorithm design. To better understand Table 2, we define $rank_{u,j}^+$ as

the proportion of users whose privacy setting for I_j is looser than u , and $rank_{u,j}^-$ as the proportion of users whose privacy setting for I_j is tighter than u . As long as we know the statistics of users' privacy setting for I_j and the adoption of u , we can compute $rank_{u,j}^+$ and $rank_{u,j}^-$.

Example 3. *Table 3 shows the statistics of birthday (I_1) privacy setting as an example. Alice allows only her friends to see her birthday.*

In Example 3, $s_1 = 4$, $rank_{Alice,1}^+ = 0.5 + 0.4 + 0.1 = 0.55$, and $rank_{Alice,1}^- = 0.1$.

In the rest of this section, we look at the insights of spacial cases in Table 2. We assume people can apply privacy setting for one type of information in Table 2. However, it is easy to extend it to multiple types of information.

Case 1 and Case 2: In fact, these two cases rarely happen in real life. We use them to demonstrate the extreme cases in PA calculations. In Case 1, Alice chooses to open her information in OSN, while others choose to hide it, which may indicate that the information is sensitive and releasing this type of information does not benefit the PIOs. In this case, Alice should get a minimum PA value due to the disclosure of sensitive information. On the other hand, if Alice chooses to hide the information, while others open it. It may means Alice is prudent when deciding to open information. Thus, Alice should get a maximum PA value in case 2.

Case 3: Without further evidence, it is difficult to interpret one's PA. Therefore, Alice has a neutral PA. In addition, if an action made by the majority, without further evidence, the action should get neither a significant negative nor positive assessment, e.g. neutral PA in TAPE.

Case 4 and Case 5: To see the insight of case 4 and case 5, we look at an example. Assume many people release birthday information to friends because they want to remind friends about their birthdays, even if they know the privacy risk.

Table 3: An example: privacy setting statistics for birthday

Privacy setting	Proportion of users adopting this privacy setting
‘everyone’	5%
‘networks’	40%
‘friends of friends’	10%
‘friends’	35%
‘self’	10%

In this case, if Alice releases her birthday, her PA should not be largely reduced. On the other hand, if Alice hides her birthday, her PA should be relatively high, since in order to gain better privacy protection, she gives up the opportunity of receiving more birthday gifts and greetings.

Case 6 and **Case 7**: Many people using tight privacy setting may imply that the information is sensitive. If Alice adopts a loose privacy setting for this type of information, her PA is low. On the other hand, if Alice adopts a tight setting, she should get a larger PA but relatively smaller than that in case 5.

We have to point out that Table 2 may not include all possible cases. For example, if Alice adopts a tight setting for birthday and a loose setting for phone number, and Bob adopts a loose setting for birthday and a tight setting for phone number, it is difficult to compare Alice’s PA with Bob’s. In such case, we need more data to make the PA evaluation more accurate. At current stage, we argue that those desirable properties in Table 2 provide a satisfying guidance for the PA algorithm design.

PA Algorithm

In our proposed PA algorithm, **individual information privacy awareness (IPA)** is calculated first. IPA is the PA value calculated from privacy setting of one type of information. Let $IPA_{u,j}$ denote the IPA of u for I_j .

$$IPA_{u,j} = \frac{1}{2}(rank_{u,j}^+ - rank_{u,j}^-) + \frac{1}{2} \quad (6)$$

where $rank_{u,j}^+$ and $rank_{u,j}^-$ are defined in Section 2.5.1. It is easy to verify that Equation 6 satisfies the desirable properties in Table 2. As an example, $IPA_{Alice,1} = 0.725$ in Example 3. Obviously, $0 \leq IPA_{u,i} \leq 1$. In fact, people can develop more sophisticated calculation to replace Equation 6, according on the implementation environment and data availability. After calculating the IPAs for all types of information, the PA of u is calculated by

$$PA_u = \frac{1}{J} \sum_{j=1}^J IPA_{u,j} \quad (7)$$

In the literature, there are some approaches proposed to evaluate similar metrics. For example, in [17], the Item Response Theory (IRT) was used for modeling “privacy concern”. In Section 2.7.1, we compare the proposed PA algorithm with IRT privacy concern model.

Privacy Trust

We propose another metric to evaluate how much a person should be trusted in terms of protecting privacy. Because this metric reflects how much one’s friends trust her/him in terms of not gossiping their information to others, it is named as **privacy trust (PT)**. In fact, this type of trust is very difficult to evaluate based on direct evidence. *First*, direct evidence is rarely available, because we cannot wait someone to commit bad behaviors (e.g. gossip others) before estimating PT. *Second*, the clues that people use to determine whether a person is trustworthy in offline social networks are usually not available in OSNs. Alternatively, indirect evidence is used to predict OSN users’ PT. Such indirect evidence can be established based on recommendations [34]. For example, Fig 6a shows a typical recommendation based trust model. If A trusts B, and B gives a recommendation saying that she/he trusts C, then A is able to develop certain level of trust to C.

In TAPE, we propose to evaluate an individual’s PT based on implicit rec-

(a) General recommendation trust model

(b) Recommendation trust model in TAPE

(c) Multiple recommendations in TAPE trust model

Figure 6: Privacy Trust model

ommendations from her/his friends. For example, an implicit recommendation for Alice from her friend Bob is established when Bob allows her to access his personal information. Moreover, if Bob has a high PA value, it implicitly tells us that Alice may be trusted not to propagate others' personal information (Fig 6b). In real life, a person working on privacy research (e.g. myself) usually has good attention to the protection of privacy, thus this person has high PA. If I choose to tell someone my personal information, it means that I trust this person not to release my personal information to others. Although such implicit recommendations have noises, it may be the best resource to compute PT in OSNs.

PT calculation in TAPE is a module whose inputs are PAs of the user's friends and trust evaluations that how much the user is trusted by friends. Let PT_u represent PT of u and G_{PT} represent PT calculation, then

$$PT_u = G_{PT}(PA_{\text{friends of } u}, T_{\text{friends}, u}) \quad (8)$$

where $T_{\text{friends}, u}$ indicates how much u is trusted by friends. Similar to PA calculation, we argue that the PT calculation should follow 3 rules.

Rule 1. *The level of PT depends on the number of positive recommendations. The recommendation from a high PA friend is considered to be a positive recommendation. The more positive recommendations a user gets, the higher her/his PT should be.*

Rule 2. *Negative recommendations should be carefully used. On the one hand, if a user with low PA trusts Alice, this should not affect Alice's PT either positively or negatively. On the other hand, if a user with high PA does not allow Alice to view personal information, it is not sufficient to indicate Alice is trustless.*

Rule 3. *Although each additional positive recommendation can increase the PT, such incremental diminishes when the number of positive recommendations is getting very large. For example, when positive recommendation number increases from 3 to 6, PT can increase a lot. However, when the number increases from 300 to 303, PT should not increase as much as the earlier case.*

PT Algorithm

In the literature, there are many trust models. We adopt a trust model using Beta function to address concatenation propagation and multi-path propagation of trust [34].

In the context of PT, the recommendation accuracy (arrow from A to B in Fig 6a) is replaced by PA of node B, and the trust value (arrow from B to C in

Fig 6a) is the implicit trust of B towards C, represented by $T_{B,C}$ in TAPE. For simplicity, in the manuscript, we set $T_{B,C}$ as a constant value, by assuming that when two users are connected in OSN, they have certain chance to see each other's personal information, but it does not necessarily mean they have already read or will read that information. In the future, when more OSN data is available, such as the nuanced privacy setting, $T_{B,C}$ can be calculated more accurately.

We use R_u^+ to denote the set of positive recommendations for u , i.e. u 's friends whose PA values are higher than a threshold (ϵ^+). The PT calculation we adopt is described as follows.

First, we estimate PT through one recommendation path

$$PT_{u,f_i} = PA_{f_i}T_{f_i,u} + (1 - PA_{f_i})(1 - T_{f_i,u}) \quad (9)$$

where $f_i \in R_u^+$ is i th high PA friend of u . Then the variance of the estimation of PT_{u,f_i} is calculated

$$\sigma_{u,f_i}^2 = \begin{cases} \frac{PT_{u,f_i}(1-PT_{u,f_i})^2}{2-PT_{u,f_i}}, & PT_{u,f_i} > 0.5 \\ \frac{PT_{u,f_i}^2(1-PT_{u,f_i})}{1+PT_{u,f_i}}, & PT_{u,f_i} < 0.5 \end{cases} \quad (10)$$

The PT is calculated using Beta trust model

$$PT_u = \frac{a}{a+b} \quad (11)$$

where $a = \sum_{i=1}^{|R_u^+|} a_i - 1$, $b = \sum_{i=1}^{|R_u^+|} b_i - 1$ and

$$a_i = PT_{u,f_i} \left(\frac{PT_{u,f_i}(1 - PT_{u,f_i})}{\sigma_{u,f_i}^2} - 1 \right) \quad (12)$$

$$b_i = (1 - PT_{u,f_i}) \left(\frac{PT_{u,f_i}(1 - PT_{u,f_i})}{\sigma_{u,f_i}^2} - 1 \right) \quad (13)$$

Calculation of NISP

PA and PT are surely two important factors that determine NISP. However, PA and PT metrics are not probability values. No theories can be used to compute NISP from PA and PT. In this work, we adopt a heuristic approach, which

Figure 7: Node ISP Calculation Diagram

estimates the NISP as a weighted average of PA and PT,

$$ISP(u) = w \cdot PA_u + (1 - w)PT_u, \quad (14)$$

where w is a weight factor between 0 and 1. Here weighted average is one of the simplest ways to combine PA and PT. In fact, People can develop more complicated calculation depending on the implementation environment and data availability. In the experiments in Section 2.7, we choose $w = 0.5$. Fig 7 shows the diagram of the NISP calculation. In the future, real human users must be involved (e.g. questionnaire) to understand the relationship among NISP, PA and PT.

2.5.2 Link Information Spreading Probability (LISP)

As discussed earlier, LISP of the link between Alice and Bob depends on whether Bob heard what Alice said. Furthermore, it depends on whether Alice has a strong tie with Bob and whether the information is interesting enough to catch Bob's attention. In the current literature, many works have investigated social ties [29, 35]. Note that the TAPE framework can accommodate any algorithms for LISP calculation, as long as the outcome of LISP calculation is a value between 0 and 1 indicating the probability of information spreading. In this manuscript, we do not propose a specific algorithm for calculating LISP. In the experiments, we adopt a constant value for LISP and focus on demonstrating the impacts of PA and PT.

2.6 Privacy Assessment and Privacy Improvement through TAPE

By evaluating NISP and LISP, and utilizing the reliability analysis method, TAPE has the ability to assess one's OSN privacy level. More importantly, based on the privacy assessment process, TAPE is able to tell people the strategies of improving privacy level.

2.6.1 Privacy Assessment

As discussed in Section 2.4, by utilizing the BDD method and adopting proper NISP and LISP algorithms, TAPE is able to evaluate privacy leakage probability from the PIO to the UD. In real life, people usually want to avoid certain personal information being viewed by multiple people, which is the reason why we define undesirable group. Without further modification, TAPE can solve multiple UD case. Given an undesirable group $UG = \{UD_1, UD_2, \dots, UD_K\}, K > 1$, the information leakage probability to UG is

$$L^{UG} = 1 - \prod_{UD_i \in UG} (1 - L^{UD_i}) \quad (15)$$

where L^i is the privacy leakage probability to UD_i . Here, privacy leakage happens if any one UD gets the information.

2.6.2 Privacy Improvement Strategies

The goal of privacy protection in TAPE is to reduce privacy risk. From a user's perspective, the most practical strategy is to block a friend to access certain personal information, also referred to as **unfriending**. In TAPE, we develop a method that can identify the friend link which contributes to the privacy leakage the most. We adopt **Birnbaum's measure (BM)** [36] to find such a friend link.

Originally, Birnbaum's measure was used to examine the sensitivity or importance of a component in reliability graph. In TAPE, we use it to evaluate the sensitivity of a friend link. Birnbaum's measure evaluates the partial derivative of

the leakage probability with respect to LISP of link c .

$$BM(c) = \frac{\partial L}{\partial ISP(c)} \quad (16)$$

For the single UD case, the detailed calculation of BM, which uses the BDD graph, can be found in [36].

We derive Birnbaum's measure for multiple UD case. By rewriting (15), we get

$$\log(1 - L^{UG}) = \sum_{UD_i \in UG} \log(1 - L^{UD_i}) \quad (17)$$

For the right-hand side,

$$\begin{aligned} & \frac{\partial}{\partial ISP(c)} \sum_{UD_i \in UG} \log(1 - L^{UD_i}) \\ &= \sum_{UD_i \in UG} \left(\frac{-1}{1 - L^{UD_i}} \frac{\partial L^{UD_i}}{\partial ISP(c)} \right) \end{aligned} \quad (18)$$

For the left-hand side:

$$\frac{\partial \log(1 - L^{UG})}{\partial ISP(c)} = \frac{-1}{1 - L^{UG}} \frac{\partial L^{UG}}{\partial ISP(c)} \quad (19)$$

Plug in together, the Birnbaum's measure of c for UG is

$$\begin{aligned} BM_{UG}(c) &= \frac{\partial L^{UG}}{\partial ISP(c)} \\ &= \sum_{UD_i \in UG} \left(\frac{1 - L^{UG}}{1 - L^{UD_i}} \frac{\partial L^{UD_i}}{\partial ISP(c)} \right) \\ &= \sum_{UD_i \in UG} \alpha_i BM_i(c) \end{aligned} \quad (20)$$

where $\alpha_i = \frac{1 - L^{UG}}{1 - L^{UD_i}}$ is the contribution weight of the i th UD, and $BM_i(c)$ is the Birnbaum's measure for c when computing the privacy leakage probability to UD_i . Finally, the unfriending strategy proposed in TAPE is to find a friend link, which is

$$c^* = \arg \max_c BM_{UG}(c) \quad (21)$$

TAPE suggests to block c^* to improve privacy level.

Table 4: Case study: Rank PA vs. IRT PA

Special Situations		Privacy setting: $s_1s_2s_3$ (0=open,1=hidden)							
		000	001	010	100	011	101	110	111
1	Percentage of users adopting this setting	70.50%	8.30%	8.92%	1.19%	8.83%	1.00%	1.10%	0.16%
	IRT z	-0.0648	0.1337	0.1337	0.3321	0.1337	0.3321	0.3321	0.5305
	Rank PA	0.4448	0.6115	0.6115	0.7782	0.6115	0.7782	0.7782	0.9448
2	Percentage of users adopting this setting	8.36%	1.18%	1.05%	0.09%	70.53%	8.68%	8.98%	1.13%
	IRT z	-0.0169	-0.0018	-0.0018	0.0133	-0.0018	0.0133	0.0133	0.0285
	Rank PA	0.3139	0.4806	0.4806	0.6472	0.4806	0.6472	0.6472	0.8139
3	Percentage of users adopting this setting	0.99%	0.17%	8.85%	0.97%	9.09%	0.93%	70.27%	8.73%
	IRT z	-0.3249	-0.1528	-0.1528	0.0194	-0.1528	0.0194	0.0194	0.1915
	Rank PA	0.1856	0.3523	0.3523	0.5189	0.3523	0.5189	0.5189	0.6856
4	Percentage of users adopting this setting	0.11%	1.20%	1.01%	9.17%	1.05%	9.16%	8.98%	69.32%
	IRT z	-0.0081	-0.0050	-0.0050	-0.0020	-0.0050	-0.0020	-0.0020	0.0010
	Rank PA	0.0569	0.2236	0.2236	0.3903	0.2236	0.3903	0.3903	0.5569

2.7 Experiment Results and Discussion

TAPE framework is implemented in Matlab, and the experiments are conducted. The reduction factor of reduced BDD is 2, i.e. $k = 2$. In PT calculation, PA threshold (ϵ^+) is 0.5 and $T_{B,C}$ is 0.7. NISP values are calculated according to Fig 7. At the beginning of this section, we first do a case study to demonstrate the calculation of PA, and then we apply TAPE to two real OSN datasets. The privacy risks and friends sensitivities are calculated. Several unfriending strategies are compared.

2.7.1 Case Study

We compare two PA algorithms. The proposed PA algorithm, referred to as Rank PA, is described in Section 2.5.1. The comparison algorithm is described in [17], referred to as IRT. Briefly speaking, this scheme calculates a metric called “privacy concern” based on privacy settings, by utilizing Item Response Theory. The goal is to estimate OSN users privacy concerns toward information sharing.

Since there is no ground truth on what should be the most “correct” value of PA, in order to demonstrate their major features, we compare these two schemes in special situations. Assume Alice has 3 types of information I_1, I_2 and I_3 , and the related privacy settings are s_1, s_2 and s_3 . The privacy setting is binary, either open (represented by 0) or hidden (represented by 1). The column index in Table 4 is the possible privacy setting. We randomly generate privacy setting data as follows. For each special situation, we first specify the proportion of each privacy configuration (e.g. ‘000’, ‘001’, etc.), and then generate the privacy configuration realities according to the distribution. 10,000 privacy configuration realities are generated for each special situation. We conduct the case studies, and the special situations we investigate are follows.

Special Situation 1 - Most users (88.91% for I_1 , 88.63% for I_2 and 89.35%

for I_3) choose to open all types of information.

Special Situation 2 - Most users (89.32%) hide type 1 information. For type 2 and type 3 information, most user (88.75% for I_2 and 88.92% for I_3) open them to public.

Special Situation 3 - Most users (89.02% for I_1 and 88.82% for I_2) hide type 1 and type 2 information, and for type 3 information 89.20% users disclose it.

Special Situation 4 - Most users (88.51% for I_1 , 88.48% for I_2 and 88.85% for I_3) hide all types of information.

The PA calculation are shown in Table 4, in which the proposed PA algorithm is referred to as “Rand PA” and the comparison scheme is referred to as “IRT z ”. We first investigate the range of each scheme. IRT z has narrow ranges for the studied situations, although the theoretical range of IRT z can be $(-\infty, \infty)$. In order to adopt IRT z in TAPE as a PA algorithm, non-trivial normalization is needed. On the other hand, the proposed Rank PA has a range from 0 to 1 as expected. In addition, the neutral value for Rank PA is 0.5, and neutral value for IRT z is 0. Then, we investigate both schemes according to desirable properties of PA in Table 2 and get follow observations.

1. The majority always get PA values close to neutral for Rank PA, i.e. “000” of special situation 1, “011” of special situation 2, “110” of special situation 3 and “111” of special situation 4. However, when IRT z is used, such majority behavior cannot be captured.
2. Special situation 1 corresponds to case 4 and case 5 in Table 2. It is seen that both Rank PA and IRT z satisfy the desirable properties of case 4 and case 5.
3. Special situation 4 corresponds to PA case 6 and PA case 7 in Table 2. Rank PA satisfies the desirable properties in PA case 6 and PA case 7. When look

at “000” in special situation 4, IRT z gives a higher value than the same privacy setting in special situation 1, which violates the desirable properties.

4. When investigating column “000” in Table 4, it is expected that the PA values from top to bottom should change from neutral to small, because the more people adopting tight privacy setting may indicate that the information is more sensitive and opening it can yield lower PA values. Rank PA has such trend, while IRT z does not.

In addition, IRT z is designed for binary privacy setting. However, real privacy setting usually has more than two options, such as in Facebook. Additionally, compare to the proposed Rank PA, IRT z also suffers from higher computational cost.

2.7.2 Datasets

We use two datasets to conduct experiments.

Dataset I contains a small number of users with detailed privacy setting. In particular, the privacy setting of 514 Facebook users in the community of University of Rhode Island (URI) are collected through a survey, and their public friends relationships are obtained by a crawler. In dataset I, there are 16 types of personal information, such as ‘email address’, ‘mobile phone number’, ‘education’ and ‘status and links’. Each type has 5 privacy setting options, including ‘everyone’, ‘networks’, ‘friends of friends’, ‘friends’ and ‘self’.

Dataset II contains a large number of users with limited privacy setting information, constructed by the authors in [37]. It contains about 957,000 Facebook users. There are 4 types of personal information, including ‘add as friend’, ‘photo’, ‘view friends’ and ‘send message’. Each type of information has 2 privacy setting options, ‘open’ and ‘hidden’.

Table 5: Datasets summary

	Dataset I	Dataset II
# of unique users	514	957,000
Average real degree	215.8	95.2
Average sampled degree	2.1	3.8
Max sampled degree	18	124
# of personal information types	16	4
Privacy setting options	‘everyone’ ‘networks’ ‘friends of friends’ ‘friends’ ‘self’	‘open’ ‘hidden’

Information of the two datasets is listed in Table 5.

2.7.3 Privacy Risk

It is well known that the reliability of data transmission can drop significantly as the distance (i.e. the number of hops) increases. In the context of privacy protection, does the privacy risk heavily depend on this distance? We study the relationship between the privacy risk and the distance from the PIO to UD.

We first randomly pick 100 nodes from dataset I and put them in the PIO set. In each round of simulation, we pick one node (without replacement) from the PIO set as the PIO, and pick another node from the network as the corresponding UD, which is no more than 6-hop away from the PIO. If the picked PIO is an isolated node (i.e. degree is 0), we skip it. For each pair of PIO and UD, we measure the distance, compute the privacy risk using TAPE, and plot the privacy risk in Fig 8. Each point represents one pair of PIO and UD. The x-axis indicates the distance between PIO and UD, and y-axis is the privacy risk. In this experiment, LISP is chosen from 0.5, 0.8, 0.9, and 0.95. We have the following observations

- As expected, when the distance increases, privacy risk has a decreasing trend.
- The privacy risk to 1-hop UDs (i.e. friends) can be greater than the LISP.

Figure 8: Privacy risk vs. PIO UD distance.

This is because Alice’s friends not only get Alice’s information from Alice directly, but also through other paths. For example, Alice’s friend Bob may not hear what Alice said, but he could get the message from Charlie who is another friend of Alice.

- When the distance is small, the privacy risk varies in a large range. The distance is not a dominating factor. The PA, PT and network topology jointly determine users’ privacy risk. A user who is 3 hops away may be more likely to obtain Alice’s person information than a user who is 2 hops away.
- As the LISP decreases, the privacy risk decreases. In the future work, incorporating the estimation of LISP will yield even a larger variation in the privacy risk values.

(a) Dataset I

(b) Dataset II

Figure 9: Histogram of proportional difference

2.7.4 The impact of PA and PT

Since the lack of “ground truth” about the real privacy risk of users, it is hard to compare TAPE with other privacy evaluation methods that consider different features of the users. Instead of comparing TAPE with a specific method, we argue that a prevalent type of privacy study in OSN only focuses on network topology. We construct a comparison method, referred to as *topology-based method*, which uses the BDD to compute the privacy risk with fixed LISP and NISP. By comparing TAPE with the topology based method, we will see whether considering PA and PT metrics reveals more information that is not captured by considering the topology alone. In the experiment, we set the LISP to be 0.5, and set the NISP of the topology based method to be the average of the NISP values when considering PA and PT.

The experiment setup is similar to that in Section 2.7.3. We construct PIO sets for both dataset I and dataset II, and each set has 100 nodes. In each round of simulation, one node is picked up (without replacement) from the PIO sets as the PIO, and another node that is 3 hops away from the PIO is picked as UD. We calculate the privacy risk using TAPE and using the topology based method. We define proportional difference as $D = \frac{V^{Topology} - V^{TAPE}}{V^{Topology}}$, where $V^{Topology}$ is the privacy risk calculated based on topology, and V^{TAPE} is the privacy risk calculated by TAPE. The histograms of D for both datasets are shown in Fig 9. It is seen that the proportional difference range is from -25% to 5%. Hence PA and PT do provide additional and useful information beyond the topology. In addition, it is seen that dataset II shows more concentrated distribution around 0, and dataset I has a wider range. It is known that, dataset II has 4 types of personal information and each type has 2 privacy setting options, while dataset I has 16 types of personal information and each type has 5 privacy setting options. We

Table 6: Impact of LISP value when comparing TAPE and the topology-based method

LISP	0.2	0.35	0.5	0.65	0.8
Avg(D)	0.0129	0.0131	0.0130	0.0127	0.0123
std(D)	0.0590	0.0595	0.0592	0.0583	0.0571

argue that the comprehensiveness of privacy setting can impact the performance of TAPE.

In the topology-based method, we set the LISP to be 0.5 and the NISP to be the average NISP in TAPE. When choosing the NISP value, we argue that the NISP setting favors the topology based method. Particularly, when PA and PT are not available, it is very difficult to choose a proper NISP value for the topology based method. By choosing the average NISP value from TAPE, we believe that it will provide a reasonable NISP estimation for the topology based method. In the rest of this section, we conduct experiment to study how much the LISP value can impact the results when comparing TAPE and the topology-based method. The experiment setup is the same as the one using dataset I earlier in this section, and we repeat it by selecting one LISP value from $\{0.2, 0.35, 0.5, 0.65, 0.8\}$ at one time. The proportional difference D is calculated. We list the statistics of D in Table 6. The proportional difference does not change when we select different LISP. However, we have to point out that smaller LISP values will give smaller privacy risk estimations, and we already observed it in Fig 8.

2.7.5 Verification of TAPE Calculation

In the previous experiments, the privacy risks are calculated from LISP and NISP using BDD as described in Section 2.4.6. In order to verify this calculation, Monte-Carlo simulations are used and the results are compared with the outputs of TAPE.

The simulation is conducted as follows. At the initial stage, a node is selected

Figure 10: TAPE vs Monte-Carlo simulation

as PIO, and the PIO owns a token, which represents one type of personal information. During the simulation stage, we divide the time into T steps. At each step, every node with a token can pass duplicates of the token to its neighbors. The probability that node A successfully passes the token duplicate to its neighbor B is $ISP(A) \cdot ISP(\langle A, B \rangle)$, where $ISP(A)$ is the NISP of A and $ISP(\langle A, B \rangle)$ is the LISP of link $\langle A, B \rangle$. After T steps, the simulation is terminated. If there is any UD that obtains a token duplicate, this simulation is marked as ‘information leakage observed’. By repeating the simulation N times, we will get N_1 ‘information leakage observed’ simulations, and the simulated privacy risk is $\frac{N_1}{N}$.

In the experiment, we randomly select 1,000 PIOs from dataset II, and those whose degrees are less than 2 are skipped. The 3-hop privacy risks are computed using TAPE. Corresponding Monte-Carlo simulations are conducted, in which $N = 100$ and $T = 5$. Fig 10 shows the results of our experiments. It is seen that the

Figure 11: Privacy improvement of unfriending

simulation justifies the calculation of TAPE, because a strong linear relationship between simulated results and TAPE results is observed. It is noticed that the slope of the curve depends on the number of simulation steps, i.e. T .

2.7.6 Sensitivity Analysis and Unfriending Strategy

Unfriending is suggested in [22, 38]. We propose an unfriending strategy based on Birnbaum’s Measure, referred to as TAPE unfriending, which evaluates the partial derivative of the leakage probability with respect to the LISP of a given friend connection. In this section, we conduct experiments to compare TAPE unfriending with 3 unfriending approaches.

1. **TAPE:** In this approach, the friend link that has the largest Birnbaum’s measure is blocked and the privacy improvement is calculated.

Table 7: Statistic summary of unfriending strategies

Strategy	Average risk reduction	std.
TAPE	0.076	0.024
Friend degree	0.052	0.039
V-Index	0.046	0.041
Random	0.039	0.039

2. **Friend Degree:** Usually, those who are popular in OSNs are considered to be the critical points in information diffusion. Therefore, we examine the privacy improvement by blocking the friend with largest degree.
3. **V-Index:** Vulnerability index was proposed by Gundecha et. al. [11], which is based on privacy setting of friends. We use this approach for unfriending, by blocking the friend with the largest V-Index.
4. **Random:** We also calculate the privacy improvement by randomly removing a friend link. This approach helps us to understand the average case when no friend sensitivity indicators are available.

The experiment setup is the same as that in Section 2.7.5. For each PIO-UD pair, we use above approaches to remove one friend link and calculate the privacy risk reductions. The experiment results are shown in Fig 11, in which the x-axis is the index of PIO-UD pair and y-axis is the privacy risk reduction. The statistic summary is shown in Table 7. We can see that TAPE gives the best performance. It is important to point out that the privacy risk reductions are calculated using the TAPE framework. It is not surprising that the Birnbaum’s measure, which is based on TAPE, performs the best. On the other hand, we show that the other unfriending strategies, which consider less information, are not as promising as the proposed TAPE framework.

2.7.7 Discussion

We define a probability based definition for privacy risk (level). Starting from the quantitative definition, the reliability evaluation method is utilized to calculate one's privacy risk in terms of information diffusion. In TAPE, PA and PT are used to capture OSN users' privacy protection behaviors. TAPE can also compute the sensitivity of one's friend links, which can assist the user to adopt unfriending strategies. TAPE can be a starting point of enhance OSN users' privacy level. Since it highly depends on data sufficiency, the OSN service providers who control the most user data could be the best candidates to implement TAPE, and their users can really benefit from it. In addition, it is expected that social links (LISP) can also impact the calculation of privacy level. Real applications should adopt an LISP algorithm while being implemented.

Privacy Leakage beyond one OSN

In TAPE, we assume that information can only be obtained through information diffusion within OSN. In practice, information diffusion is a much more complex process. There are several scenarios of information diffusion in social networks.

1. **Cross-OSN diffusion:** People can be active in multiple OSN platforms. For example, Alice is a friend of Bob on Twitter. She sees news about Bob on Twitter, and then she posts some words about the news on Facebook.
2. **Offline diffusion:** This is the traditional way we spread information through face-to-face conversation, phone calls etc.
3. **Online-Offline diffusion:** Information is propagated through both online and offline channels. This is the most common way we spread information in the information era.

Whereas scenario 2 is well studied in social science, scenarios 1 and 3 are challenging. In all the three scenarios, the concepts of privacy awareness and privacy trust are still valid. They have a great potential to be adopted in these scenarios and contribute to a broader study on personal privacy leakage in a hybrid online-offline world.

2.8 Conclusion and Future Work

In this manuscript, we present a TAPE framework for the quantitative evaluation of users' privacy risk in OSNs. Mathematical tools (e.g. statistics, modeling techniques) are used to process online social network data, and signal processing tools are utilized in this work. The concepts of privacy awareness and privacy trust are introduced. Simulations are performed to illustrate the computation of privacy leakage probability, as well as to demonstrate that TAPE can capture useful information which was not captured previously. Several unfriending strategies are compared with the Birnbaum's method of TAPE, and TAPE gives the best performance. More importantly, TAPE sets up the stage for utilizing reliability analysis, which is a well-developed field, to solve privacy risk analysis problems. Besides BDD, other tools such as sensitivity analysis can surely benefit privacy research.

Future work includes developing better PA and PT algorithms, implementation of TAPE as a Facebook application and performing real user testing.

List of References

- [1] d. m. boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.
- [2] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*, Alexandria, VA, USA, 2005.

- [3] B. Krishnamurthy and C. E. Wills, “On the leakage of personally identifiable information via online social networks,” in *Proceedings of the 2nd ACM workshop on Online social networks*, Barcelona, Spain, 2009.
- [4] S. Livingstone, “Taking risky opportunities in youthful content creation: teenagers’ use of social networking sites for intimacy, privacy and self-expression,” *New media & society*, vol. 10, no. 3, pp. 393–411, 2008.
- [5] EurActiv. “Social networks put careers at risk, survey finds.” 2010. [Online]. Available: <http://www.euractiv.com/Social-networks-careers-risk>
- [6] S. Strauß and M. Nentwich, “Social network sites, privacy and the blurring boundary between public and private spaces,” *Science and Public Policy*, vol. 40, no. 6, pp. 724–732, 2013.
- [7] S. E. Schechter, “Computer security strength and risk: A quantitative approach,” Ph.D. dissertation, Harvard University, Cambridge, MA, USA, 2004.
- [8] G. Stoneburner, A. Goguen, and A. Feringa, “Risk management guide for information technology systems,” *Nist special publication*, vol. 800, no. 30, 2002.
- [9] H. In, Y.-G. Kim, T. Lee, C.-J. Moon, Y. Jung, and I. Kim, “A security risk analysis model for information systems,” in *Systems Modeling and Simulation: Theory and Applications*. Springer Berlin Heidelberg, 2005, vol. 3398, pp. 505–513.
- [10] M. Hamdi and N. Boudriga, “Computer and network security risk management: theory, challenges, and countermeasures,” *International Journal of Communication Systems*, vol. 18, no. 8, pp. 763–793, 2005.
- [11] P. Gundecha, G. Barbier, and H. Liu, “Exploiting vulnerability to secure user privacy on a social networking site,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, San Diego, CA, USA, 2011.
- [12] L. Fang and K. LeFevre, “Privacy wizards for social networking sites,” in *Proceedings of the 19th international conference on World Wide Web*, Raleigh, NC, USA, 2010.
- [13] R. Baden, A. Bender, N. Spring, B. Bhattacharjee, and D. Starin, “Persona: an online social network with user-defined privacy,” in *Proceedings of the ACM SIGCOMM 2009 conference on Data communication*, Barcelona, Spain, 2009.
- [14] A. Felt and D. Evans, “Privacy protection for social networking APIs,” in *Web 2.0 Security and Privacy*, Oakland, CA, USA, 2008.

- [15] A. Jeckmans, M. Beye, Z. Erkin, P. Hartel, R. Lagendijk, and Q. Tang, "Privacy in recommender systems," in *Social Media Retrieval. Computer Communications and Networks*. Springer Verlag, London, 2013, pp. 263–281.
- [16] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *Proceedings of the 18th international conference on World wide web*, Madrid, Spain, 2009.
- [17] S. Guo and K. Chen, "Mining privacy settings to find optimal privacy-utility tradeoffs for social network services," in *International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing*, 2012.
- [18] S. Guha, K. Tang, and P. Francis, "NOYB: privacy in online social networks," in *Proceedings of the first workshop on Online social networks*, Seattle, WA, USA, 2008.
- [19] R. De Wolf, R. Heyman, and J. Pierson, "Privacy by design through a social requirements analysis of social network sites from a user perspective," in *European Data Protection: Coming of Age*. Springer Netherlands, 2013, pp. 241–265.
- [20] Z. Erkin, T. Veugen, and R. Lagendijk, "Generating private recommendations in a social trust network," in *International Conference on Computational Aspects of Social Networks (CASON '11)*, Salamanca, Spain, 2011.
- [21] S. Alim, D. Neagu, and M. Ridley, "A vulnerability evaluation framework for online social network profiles: axioms and propositions," *International Journal of Internet Technology and Secured Transactions*, vol. 4, no. 2, pp. 178–206, 2012.
- [22] R. Abdulrahman, S. Alim, D. Neagu, and M. Ridley, "Algorithms for data retrieval from online social network graphs," in *IEEE 10th International Conference on Computer and Information Technology*, Bradford, UK, 2010.
- [23] R. Abdulrahman, S. Alim, D. Neagu, D. R. W. Holton, and M. Ridley, "Multi agent system approach for vulnerability analysis of online social network profiles over time," *International Journal of Knowledge and Web Intelligence*, vol. 3, no. 3, pp. 256–286, 2012.
- [24] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proceedings of the 13th international conference on World Wide Web*, New York, NY, USA, 2004.
- [25] E. Adar and L. Adamic, "Tracking information epidemics in blogspace," in *IEEE/WIC/ACM International Conference on Web Intelligence*, Compigne, France, 2005.

- [26] F. Wang, H. Wang, and K. Xu, “Diffusive logistic model towards predicting information diffusion in online social networks,” in *32nd International Conference on Distributed Computing Systems Workshops*, Macau, China, 2012.
- [27] M. Z. Shafiq and A. X. Liu, “Modeling morphology of social network cascades,” *Computing Research Repository*, vol. abs/1302.2376, 2013.
- [28] E. Sadikov, M. Medina, J. Leskovec, and H. Garcia-Molina, “Correcting for missing data in information cascades,” in *Proceedings of the fourth ACM international conference on Web search and data mining*, Hong Kong, China, 2011.
- [29] J. Zhao, J. Wu, X. Feng, H. Xiong, and K. Xu, “Information propagation in online social networks: a tie-strength perspective,” *Knowledge and Information Systems*, vol. 32, no. 3, pp. 589–608, 2012.
- [30] Y.-G. Kim and J. Lim, “Quantitative risk analysis and evaluation in information systems: A case study,” in *Computational Science ICCS 2007*. Springer Berlin Heidelberg, 2007, vol. 4489, pp. 1040–1047.
- [31] X. Zang, H. Sun, and K. S. Trivedi, “A bdd-based algorithm for reliability graph analysis,” *Technical report, Department of Electrical Engineering, Duke University*, 2000.
- [32] C. Wang, L. Xing, V. Vokkarane, and Y. Sun, “Reliability analysis of wireless sensor networks using different network topology characteristics,” in *International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering*, Chengdu, China, 2012.
- [33] C. Wang, L. Xing, V. M. Vokkarane, and Y. Sun, “Manycast and anycast-based infrastructure communication reliability for wireless sensor networks,” in *The 18th ISSAT International Conference on Reliability and Quality in Design*, Boston, MA, USA, 2012.
- [34] Y. Sun, Z. Han, W. Yu, and K. Liu, “A trust evaluation framework in distributed networks: Vulnerability analysis and defense against attacks,” in *25th IEEE International Conference on Computer Communications. Proceedings*, Barcelona, Spain, 2006.
- [35] R. Xiang, J. Neville, and M. Rogati, “Modeling relationship strength in online social networks,” in *Proceedings of the 19th international conference on World wide web*, Raleigh, NC, USA, 2010.
- [36] J. Andrews and S. Beeson, “Birnbbaum’s measure of component importance for noncoherent systems,” *IEEE Transactions on Reliability*, vol. 52, no. 2, pp. 213–219, 2003.

- [37] M. Kurant, M. Gjoka, C. T. Butts, and A. Markopoulou, “Walking on a graph with a magnifying glass: Stratified sampling via weighted random walks,” in *Proceedings of ACM SIGMETRICS*, San Jose, CA, 2011.
- [38] C. Sibona and S. Walczak, “Unfriending on facebook: Friend request and online/offline behavior analysis,” in *44th Hawaii International Conference on System Sciences*, Kauai, HI, USA, 2011.

CHAPTER 3

“Equal Rating Opportunity Analysis for Detecting Review Manipulation”

by

Yongbo Zeng¹, Yihai Zhu², and Yan (Lindsay) Sun³

is in preparing for submission to IEEE Transactions on Information Forensics
and Security

An earlier and shorter version is published in the IEEE China Summit and
International Conference on Signal and Information Processing (ChinaSIP 2015)

¹PhD Candidate, Department of Electrical, Computer & Biomedical Engineering, The University of Rhode Island, Kingston, RI 02881. Email: yongbozeng@ele.uri.edu

²PhD, Tableau Software, Email: yihaizhu@gmail.com

³Professor, Department of Electrical, Computer & Biomedical Engineering, The University of Rhode Island, Kingston, RI 02881. Email: yansun@ele.uri.edu

3.1 Abstract

Online market has been growing for many years. There have been huge profits in online markets. Online product/service review plays an important role when customers are shopping on the Internet. The purchasing decisions dramatically depend on the product/service reviews. Sellers can benefit from creating fake reviews to boost their products' reputation or bad-mouth their competitors' products. Fake reviews can seriously affect both buyers and sellers. In this manuscript, we introduce a novel angle to detect fake reviews, which is called Equal Rating Opportunity (ERO) principle. Based on ERO principle, we propose a fake review detection method – ERO analysis, which is able to detect fraudulence signals based missed by existing approaches. Experiments based on Amazon product reviews are conducted, in which the applicability of review features for ERO analysis is studied, two common fake review detection methods are compared with the proposed ERO analysis, and expert reviews are employed to evaluate the performance of ERO analysis. The ERO analysis provides a new angle of fake review detection, which is important when the fake reviews are diverse.

3.2 Introduction

Online product/service reviews are created by customers who have experiences of using the product/service, such as Amazon product review, Yelp restaurant review and TripAdvisor travel review. A piece of review often consists of a rating value that represents the user's overall satisfaction, and a text that describes the experience of purchasing and using the product/service. Online review system, also referred to as *online reputation system*, allows users to post reviews for products/services. The reviews of a single product/service are often combined to an overall reputation score, such as the average number of stars in Amazon. Product/service reviews have shown huge impacts to online sales. It is reported that a

one-star rating increase can bring a 5-9% increase of revenue to online sellers [1]. On the one hand, the average rating can affect products' display order on the site. Products with higher average ratings are often put in front of the list and thus can get more visibilities. On the other hand, customers are relying on the average rating when they are screening products. It is reported 74% of consumers rely on product reviews to guide their purchases [2].

However, online reviews may be manipulated. It is reported that sellers at the online marketplace boost their reputation by trading with collaborators [4], and firms post fake reviews to praise their own products or bad-mouth their competitors' [5]. Review manipulation can inflate or deflate products' reputation scores, crash users' confidence in online reputation systems, and eventually undermine reputation-centric online businesses leading to economic loss. Furthermore, there are some situations, in which the review manipulation is even more damaging. For example, *Black Friday* shoppers heavily relies on online reviews, because they have to make rush decisions for the products they are not familiar with in order to take advantage of these quickly expiring 'unusual' discount. Another example is online reputation of hotels and restaurants. The consumers, who are misled by manipulated hotel ratings, cannot be easily refunded after they purchase these services.

In the literature, researchers propose methods to protect reputation systems from several angles, such as 1) increasing the cost of registering multiple user accounts [6], 2) endogenous discounting of fake reviews by analyzing the statistic features of reviews [7], 3) exogenous discounting of dishonest ratings by introducing metrics of users' reputation [7–9], and 4) studying correlation between users and reviews to detect fake reviews [10, 11]. In this manuscript, we propose a new angle to detect products with fake reviews, which is called Equal Rating Opportunity

(ERO) analysis. This method roughly belongs to category 2 and 4.

Fake review detection, also referred to spam review detection in some literature, is an urgent yet under investigation task. Unlike other types of spam, such as spam email, fake review is much harder to detect. The main reason is that fake reviewers can easily pretend to be honest. It is hard for a human user to recognize them. Some work started the research of fake review detection many years ago. For example, the opinion spam detection was proposed by Jindal and Liu in [12], in which the authors built a classifier to detect review duplications. There is also an example in the commercial field. Yelp has been using a review filter to hide certain suspicious reviews since 2005 [13, 14]. We argue that fake review detection has following challenges, which are the major obstacles of this work.

1. Fake reviewers' behaviors may be hard to capture, and the fake reviews' structures are diverse and usually unknown. For example, in order to successfully mislead review readers, fake reviewers can make their writing styles and review habits look very similar to honest reviewers. Without knowing the structures of fake reviews, it is difficult to distinguish fake ones from genuine reviews.
2. Fake reviewers can learn from the detection strategy to avoid detection. Therefore, fake review detection approaches have to be robust. For example, some existing work evaluate and assign trust scores to reviewers. Fake reviewers can post genuine reviews to boost their trust score before they creating fake reviews.
3. There is no ground truth whether a review is faked or not. By reading the review text alone, we usually do not have enough clues to tell dishonest ones from honest reviews.

These are the major challenges that we think make simple behavioral heuristics insufficient. To detect such elusive fake reviews, we need to consider more clues. In this work, we introduce the ERO principle. The ERO principle tells us that the sentiment of the review, such as the rating value, should not depend on certain review features. If dependency or correlation is observed, it is highly suspicious that the product contains fake reviews. Based on the ERO principle, ERO analysis is proposed. ERO analysis has two major advantages.

- It does not require the cooperation of online reputation system owners. In particular, many existing algorithms need to use a large amount of data, which makes them impractical unless the reputation system owners (e.g. Amazon and Yelp) implement these algorithms. Our approach, however, can be implemented by a third party, who only needs to crawl a very small amount of data to train the detection parameters and perform the ERO analysis. Therefore, the proposed method is a low-cost solution, yields independent opinions, and leads to practical implementations.
- ERO analysis is a new direction of fake review detection. It is compatible with most of existing algorithms, which exam ratings and reviewers from more traditional angles. ERO has a potential to find the fraudulence signals, which were previously missed by existing approaches.

As a summary, the **contributions** of this work include:

1. The ERO principle is introduced. It is a new perspective how we can investigate product/service reviews;
2. A new categorization of review features is introduced, and the criterion of ERO feature selection is studied;

3. The ERO analysis is proposed and real data evaluation is performed. Particularly, expert reviews are employed to evaluate the detection accuracy.

This manuscript is organized as follows. Related work is discussed in Section 3.3. The ERO principle and ERO analysis are presented in Section 3.4, followed by the discussion of ERO feature selection in Section 3.5. Experiments and results are shown in Section 3.6.

3.3 Related Work

In order to protect online reputation systems, researchers propose many protection schemes, which can be roughly put into 4 categories. The **first** category is increasing the cost of getting multiple user accounts by binding user IDs with IP addresses [6]. The **second** category is endogenous discounting of fake reviews [7]. Dishonest ratings are directly differentiated from normal ratings based on the statistic features of the rating values. In a Beta-function based approach [15], a user is determined as a malicious user if the estimated reputation of the product rated by him/her lies outside q and $1 - q$ quantile of his/her underlying rating distribution. An entropy based approach is proposed in [16]. The **third** category is exogenous discounting of dishonest ratings. Users are assigned trust scores based on their review history, and the quality of their reviews are discounted according to their trust scores. In [9], a user's trust is obtained by cumulating his/her neighbors' beliefs through belief theory. The **fourth** category is studying correlation between users and reviews to detect dishonest ratings [10, 11]. The proposed scheme has both category 2 and category 4 features, and the detection algorithm is from a new angle.

Fake reviews can undoubtedly reduce the quality of reviews. They may even mislead users to make wrong purchase decisions. Therefore, there is a great demand to detect fake reviews thoroughly on the reputation system. There are three

ways that can achieve the detection of fake reviews. The first way is directly detecting fake reviews, primarily based on the features of review. For example, the standard word and part-of-speech n-gram features are used in [17] to identify fake reviews. Another major feature of fake review detection is review text. Some approaches rely on the identification of duplicated or near-duplicated text occurring in multiple reviews [18, 19]. The second way is dishonest reviewer detection. For example, a graph-based method is used to find fake store reviewers [20], frequency pattern mining is employed to find groups of reviewers who frequently write reviews together [21], and user correlations are analyzed to identify fake reviewer groups [10]. The third way is victim product detection. A product is considered to be a victim if positive fake reviews or negative fake reviews are detected. This type of detection often assumes norms of review statistics. For example, in [10], the CUSUM approach is used, and the average rating is assumed not to jump/drop dramatically in a short period of time. In [22], the changes of mean value and changes of arrival rate are considered to be signals of fake reviews. The proposed ERO analysis belongs to victim product detection.

Many research results did not turn into practical systems. This is probably because of the potential liability concerns of major e-commerce companies, as well as the gap between research and practical constraints. Without the support from the e-commerce companies (i.e., reputation system owners), the algorithms can only rely on a limited amount of data. This is one of the major hurdles. Currently, there are only a few existing online systems providing review analysis services. For example, there is a website called “ReviewPro” [23], whose major business is to provide professional suggestions to hotel owners. By analyzing the customers’ reviews on a hotel, ReviewPro can provide analytical reports with “strategies” to climb TripAdvisor rankings and earn 5-star reviews. Another practical system

is “TrustYou” [24], which provides review analysis services on hotels. For hotel owners, it provides service to market the reputation and increase businesses. For individual users, it provides service to analyze the hotel’s quality, by summarizing online reviews and generating a trust score for the hotel. What we propose in this work is fundamentally different from these existing services. First, our work focus on detecting review manipulation, instead of finding patterns for reputation promotion purpose. Second, our work can provide on-demand real time service, whereas ReviewPro and TrustYou can only offer analysis of a pre-determined list of hotels. That is, our algorithm is so effective that it can detects fraudulence signals based on the small amount of data crawled in real time.

3.4 ERO Principle and ERO Analysis

A product is considered to be a victim if it has fake reviews, either positive, e.g. fake 5 star reviews, or negative, e.g. fake 1 star reviews. Victim product detection aims to determine if a product contains fake reviews. In this section, ERO principle and ERO analysis are introduced. To help the readers understand the underlying thoughts of ERO principle, we first give a brief overview of a consistency analysis method, which is introduced in [25].

3.4.1 Consistency Analysis

Consistency analysis is a victim product detection technology. The unusual jump/drop of review statistics is considered inconsistent. Consistency detectors are based on the fact that in order to perform an effective review manipulation, the fake reviews must cause large enough change in the review statistics, such as the average rating. In the literature, there are several approaches to detect the inconsistency of reviews [22, 25]. In this subsection, we give a brief overview of the one proposed in [25] called CUSUM.

First of all, the notations used in this manuscript are defined as follows.

- p_i is the product with product id i .
- $r_{i,n}$ is the n th review of p_i , where the reviews are sorted by posting time from old to new. $n = 1, 2, \dots, N$.
- $r_{i,n} = \{x_{i,n}, f_{i,n}^{(1)}, f_{i,n}^{(2)}, \dots, f_{i,n}^{(K)}\}$, where $x_{i,n}$ is the rating, and $f_{i,n}^{(K)}$ is the observation of the k th ERO feature. $k = 1, 2, \dots, K$
- Vector notations are also used. $\mathbf{r}_i$ is the review vector, $\mathbf{x}_i$ is the rating vector, and $\mathbf{f}_i^{(k)}$ is the ERO feature observation vector.

Let μ_i be the true rating of p_i and ν be trigger threshold, if $\bar{\mathbf{x}}_i > \mu_i + \nu$ or $\bar{\mathbf{x}}_i < \mu_i - \nu$, then jump/drop of average rating is observed, where $\bar{\mathbf{x}}_i = \frac{1}{N} \sum_{n=1}^N x_{i,n}$ is the average rating. The detection functions are defined as follows.

$$\begin{cases} g_{i,n}^+ = \max(g_{i,n-1}^+ + x_{i,n} - \mu_i - \nu/2, 0) \\ g_{i,n}^- = \max(g_{i,n-1}^- - x_{i,n} + \mu_i - \nu/2, 0) \end{cases} \quad (22)$$

where $g_{i,n}^+$ means the positive changes, $g_{i,n}^-$ means the negative changes, and $g_{i,0}^+ = 0$, $g_{i,0}^- = 0$ for initialization. *Rating inconsistency* is observed when g_n^+ or g_n^- exceeds the threshold $\bar{h}$.

In order to measure the degree of inconsistency, another metric, Percentage of Change Interval (PCI), is defined as

$$PCI(\bar{h}) = \frac{N_D}{N} \quad (23)$$

where N_D is the number of $g_{i,n}^+$ or $g_{i,n}^-$ points exceeds $\bar{h}$. Obviously, PCI value depends on the selection of threshold.

It is pointed that a uniform threshold for all products is not applicable, and heterogeneous thresholds should be used [25]. Briefly speaking, the threshold depends on $PCI(\bar{h}_0)$, where $\bar{h}_0$ is a predefined minimum threshold. Smaller $PCI(\bar{h}_0)$

gives smaller threshold, and vice versa. In this manuscript, we use the PCI value $PCI(\bar{h})$ as the evaluation of rating inconsistency. $PCI(\bar{h}) = 0$ means the ratings are consistent, $PCI(\bar{h}) = 1$ means strong inconsistency is observed. More detailed discussion on PCI can be found in [25]

3.4.2 ERO Principle

The consistency detection, which is based on the statistics of reviews, can only be used to find products that are suspected to be under review manipulation, but is lack of the capability to accurately detect such manipulation. This is because the normal rating can change without any manipulation. For example, when a restaurant changes the chief, a seller changes his/her attitude toward consumer complaints, and the manufacturer fixes a defect of the product, the ratings for the restaurant/seller/product could change. The rating is also related to price. Consumers tend to be more tolerant if they purchase deeply discounted products. If the price changes dramatically, the ratings may change. Therefore, after the consistency analysis gives us a set of suspicious products, we must apply a more informative analysis to confirm the review manipulation.

We are inspired by the *Equal Employment Opportunity Policy*, adopted by many employers. One example of such policy statement is as follows.

“All employment decisions at the company are based on business needs, job requirements and individual qualifications, without regard to race, color, religion or belief, national,”

We introduce the **Equal Rating Opportunity (ERO) Principle**, as follow.

ERO Principle. *“The normal ratings should be primarily based on the quality of the product or service, without regard to whether the review is posted on weekdays or weekend, posted during daytime or night time,”*

This principle is based on the idea that the fake reviewers may maintain a lot

of user accounts and review templates, and if they perform review manipulation on a product, it can change the distribution of some features. It is also possible that some fake reviewers are not working full time, they may focus on writing fake reviews during their spare time, such as evenings and weekends. This type of unusual correlation is called *fraudulence signal*. For example, if the fake positive reviews are posted on specific days, such as weekends, it will increase the correlation between the ‘day of week’ feature and rating value. If the fake reviewers randomly select user accounts from a large account pool and post fake reviews on a random day, the fraudulence signal can also be detected as long as an appropriate feature is used, such as *whether the reviewer purchased the product* and *reviewer city*.

3.4.3 ERO Feature

Review features applied to ERO principle are called **ERO feature**. A review often has multiple features, such as rating value, review text, review date, reviewer’s name and reviewer’s reviewing history. Obviously, some review features may not be applicable for ERO principle. For example, people who give 5 star reviews in Amazon tend to write short text, and those who give 1 star and 2 star reviews are more likely to write long text to express their complaint. Therefore, the text length being correlated to the rating value cannot be considered as a violation of ERO principle. This example is also observed in Section 3.5.

If we can find a review feature that statistically cannot be correlated to rating value, then the feature is called *ERO feature*. ERO principle holds upon ERO features. One example is the ‘day of week’ feature. The percentage of negative review and the percentage of positive review cannot significantly vary when we look at on which day of a week the reviews were posted. If the percentage of positive reviews in weekends are found much higher than on weekdays, it is highly suspicious that the product suffers fake reviews. Some other ERO features examples are ‘time

in a day’ and ‘verified’. The detailed discussion of ERO feature and ERO feature selection is presented in Section 3.5.

3.4.4 ERO Analysis

Based on ERO principle, we propose ERO analysis for fake review detection. ERO analysis takes the rating value vector $\mathbf{x}_i$ and the ERO feature observation vector $\mathbf{f}_i^{(k)}$ as inputs, and outputs the ERO value as well as the detection results, i.e. *suspicious* or *normal*. Let $E_i^{(k)}$ denote the ERO value upon feature $\mathbf{f}_i^{(k)}$. We employ the Pearson correlation coefficient as the ERO calculation.

$$E_i^{(k)} = \frac{\text{cov}(\mathbf{x}_i, \mathbf{f}_i^{(k)})}{\sigma_{\mathbf{x}_i} \sigma_{\mathbf{f}_i^{(k)}}} \quad (24)$$

where $\text{cov}(\mathbf{x}_i, \mathbf{f}_i^{(k)})$ is the covariance of $\mathbf{x}_i$ and $\mathbf{f}_i^{(k)}$, $\sigma_{\mathbf{x}_i}$ is the standard deviation of $\mathbf{x}_i$, and $\sigma_{\mathbf{f}_i^{(k)}}$ is the standard deviation of $\mathbf{f}_i^{(k)}$. 0 ERO value means no correlation. 1 or -1 ERO value means absolutely correlated. Generally, the larger the absolute ERO value is, the stronger the correlation is.

A nonzero ERO value is not sufficient to make decision, because the ERO value can be noisy. First, ERO value is an estimation of the Pearson correlation coefficient. The estimation error depends on the number of samples. Second, honest reviewers can have bias on their reviews, based on their preferences for example. Some reviewers may be hypercritical and give low rating just because of tiny problems; Some other reviewers may tolerate tiny problems and still speak highly of the product. In this work, the *boxplot* [26] technology is used to determine the detection thresholds. Briefly speaking, the thresholds depend on the observed noise.

For a given ERO feature $\mathbf{f}_i^{(k)}$, we have M products to calculate the detection thresholds. First, the ERO value of each product, $E_{d1}^{(k)}, E_{d2}^{(k)}, \dots, E_{dM}^{(k)}$, is calculated by Equation 24. Next, the *first quartile*, denoted by q_1 , and the *third quartile*,

denoted by q_3 , are calculated respectively. Finally, the detection thresholds are calculated as

$$\begin{cases} \theta_{upper} = q_3 + w(q_3 - q_1) \\ \theta_{lower} = q_1 - w(q_3 - q_1) \end{cases} \quad (25)$$

where w is the sensitivity factor. Smaller w makes the detector easier to alarm and thus can achieve higher detection rate as well as higher false alarm rate. $(\theta_{upper}, \theta_{lower})$ is called safe interval. ERO values within the safe interval are considered normal, otherwise suspicious. In other words, if $E_i^{(k)} < \theta_{lower}$ or $E_i^{(k)} > \theta_{upper}$, then the ERO analysis upon feature $\mathbf{f}_i^{(k)}$ alarms that p_i is *highly suspicious suffering fake reviews*.

There can be multiple ERO features, which means $K > 1$. Different detection thresholds are used for different ERO features. Meanwhile, multiple ERO features provide multiple detection results. In this manuscript, we consider the product to be suspicious when at least one ERO feature alarms. In practice, the detection fusion method can be more sophisticated, depending on the number of ERO features and the required detection rate or false alarm rate.

3.4.5 Case Study

In this section, we use 4 real products to demonstrate the proposed ERO principle and ERO analysis. The review data are collected from Amazon (see Section 3.6.2 for more details).

We use two ERO features – ‘verified’ and ‘day of week’. Briefly, a review is said ‘verified’ if the reviewer purchased the product. The ‘day of week’ feature means on which day of a week the review was posted. We select four typical products. According to ERO analysis, two of them are considered normal and the other two are determined suspicious. The review details are shown in Fig 12.

In Fig 12a, we observed that the percentage of non-verified reviews is not related to the rating, for which the ERO value is 0.04. Similarly, in Fig 12c, the

(a) Normal case, 'verified' feature

(b) Suspicious case, 'verified' feature

(c) Normal case, 'day of week' feature

(d) Suspicious case, 'day of week' feature

Figure 12: ERO principle examples

percentages of positive review (4 star and 5 star) and negative review (1 star and 2 star) are not related to 'day of week' feature, for which the ERO value is -0.03. Therefore, those two products are determined to be *normal* by ERO analysis. Conversely, in Fig 12b, the 'verified' feature shows certain correlation with rating value, for which the ERO value is 0.42. In Fig 12d, the percentage of negative reviews on Sundays is much different from on other days, for which the ERO value is 0.33. Since relatively strong correlations are observed, those two products are determined to be *suspicious*.

3.5 ERO Feature Selection

As we discussed earlier in Section 3.4.3, a reviews often has multiple features. Some of them are applicable for ERO principle, and some other are not. In this section, we discuss the strategies how ERO features are selected.

3.5.1 Review Feature Categorization

In some literatures, review features are categorized into three types – product specific feature, review specific feature and reviewer specific feature, based on the feature origins [17, 27]. Product specific feature, also referred to as product meta-information, describes the product’s properties, such as price, discount rate and category. Review specific feature describes the review’s characteristics, such as review text, number of votes and review date. Reviewer specific feature describes the reviewer’s profile and history, such as number of reviews posted by the reviewer, reviewer ranking and number of votes.

In the context of ERO analysis, since the analysis is conducted for specific product, the product specific feature is not used. ERO features are selected from the review specific feature and the reviewer specific feature categories.

When considering the features’ applicability of ERO analysis, we categorize review features into three groups rather than the ones mentioned earlier.

- **Persistent Feature** is defined to be the features being determined when the review was posted and will never change. For example, the review date was determined when a review was posted, and it cannot change again. Another example is ‘verified’ feature. Whether a review is verified has to be determined at the time when the review was posted based on the reviewer’s purchase history.
- **Acquired Feature** is defined to be the features acquired after the review

was posted. For example, when people think a review is useful, they will vote it as helpful and thereby increase the number of helpful votes and total votes. The number of votes may keep changing as long as it is accessible.

- **Accumulative Feature** is defined to be the features that had already existed before the review was posted and will keep changing. For example, the number of lifetime reviews of a reviewer will keep increasing as long as the reviewer keeps posting reviews.

We argue that the ERO features should be selected from the Persistent category. The Acquired and Accumulative categories are usually not suitable for ERO analysis. We summarize two criterions that can guide ERO feature selection.

The first criterion is *causality*. Recall that the ERO principle tells us that the rating value should depend on the product quality and purchase experience but other factors, i.e. ERO features. In other words, it is impossible to predict the rating value based on uncorrelated features. For example, when considering whether the user purchased the product, reviews can be divided into two groups – verified and non-verified. The ‘verified’ information cannot help us predict the rating value. In contrast, we observed that negative reviews tend to have longer review text (see Section 3.5.2). Therefore, given the text length of a review, we are able to predict the rating value, and the prediction should be more accurate than guessing. Theoretically, causation implies correlation, but correlation does not necessarily imply causation. Therefore, an ERO feature should not have causality with rating value directly or indirectly.

The second criterion is *stability*. If a review feature’s observations are changed after reviews were posted, then the feature is said unstable. An unstable feature is not suitable for ERO analysis, because the observations depend on some other factors. For example, there are two similar reviews – R_1 and R_2 , and they are

Table 8: Review features

Name	Type	Description
Total votes	Non-negative integer	The number of votes given by review readers.
Day of week	1 to 7	On which day of week the review was posted. It starts from Sunday (value=1).
Content length	Non-negative integer	The length of the review text, counting by characters.
Reviewer’s number of lifetime reviews	Positive integer	The number of reviews posted by the reviewer as of the time of data collection.
Reviewer ranking	Positive integer	Reviewer ranking in Amazon system. Smaller ranking value means higher ranking.
Verified	Boolean	Whether the reviewer purchased the product when they posting the review.

considered to be the same helpful. However, R_1 was posted one year earlier than R_2 . The number of helpful votes of R_1 could be much higher than R_2 . In addition, some unstable features can promote other unstable features. For example, the reviewer’s number of lifetime reviews, reviewer ranking and number of helpful votes may affect the review display order. The display order can also affect the number of helpful votes and thereby reviewer ranking. The correlation of unstable features usually only reveals the fraudulence signal, but also some implicit relationship. Therefore, the Acquired and Accumulative categories are not suitable for ERO analysis.

In summary, a review feature is **not** applicable for ERO analysis, if 1) it is not Persistent, 2) causality observed, or 3) it is unstable.

3.5.2 Feature Comparison

In this subsection, we use real data to demonstrate the feature selection process. The data were collected from Amazon (refer to Section 3.6.2). There are total 416 products randomly selected from the dataset. The review features investigated in this section are listed in Table 9. We calculate the ERO values of all

Figure 13: Review features

those features for the 416 products. The results is plotted in Fig 13.

Total Votes

‘Total votes’ feature means the number of votes the review got. From Fig 13, we observe that it has negative correlation with the rating value. In other words, negative reviews (1 star and 2 star) potentially get more votes than positive ones. When looking at the reviews, we find that negative reviews often have longer and more informative texts while positive reviews often have shorter texts. The possible reason is that when people are reading reviews, they tend to give votes to the reviews with longer texts.

Day of Week

From Fig 13, we can see the irrelevance between ‘day of week’ feature and rating value. Statistically, the numbers of reviews on a day of week may not be identical, but the ratio between negative and positive review numbers can not

significantly vary. Therefore, small ERO values are observed.

Content Length

The ‘content length’, also called text length, has negative correlation with the rating value. It is likely that customers with negative experience have more to complain, and therefore their review texts are longer. Note that this observation also confirms the one we get for the ‘total votes’ feature.

Reviewer’s Number of Lifetime Reviews

When investigating the number of lifetime reviews that the reviewer has posted, we find that reviewers with more reviews tend to post positive reviews. In Fig 13, a positive correlation is observed. It is likely that experienced reviewers are more tolerant to negative experience.

Reviewer Ranking

The reviewer ranking has negative correlation with the rating value, which means higher ranked reviewers tend to give positive reviews. It also partially confirms the observation of ‘reviewer’s number of lifetime reviews’, since in Amazon system, the reviewer ranking has certain relationship to the number of lifetime reviews.

Verified

A review is verified if the reviewer purchased the product before he/she posting the review. In Fig 13, there is no strong correlation observed for ‘verified’ feature.

Summary

The average correlations are calculated and shown in Table 9. In summary, there are two features applicable for ERO analysis, including ‘day of week’ and ‘verified’. There are correlations observed for the other features, and some of

Table 9: Review features summary

Name	Mean correlation	Std. correlation	Category	Causality	Stability
Total votes	-0.0844	0.1029	Acquired	Probably	No
Day of week	-0.0067	0.0642	Persistent	No	Yes
Content length	-0.1245	0.1101	Persistent	Probably	Yes
Reviewer number of reviews	0.0836	0.0693	Accumulative	Probably	No
Reviewer ranking	-0.0716	0.0895	Accumulative	Probably	No
Verified	0.0080	0.0604	Persistent	No	Yes

them have causality observed with rating value, such as ‘total votes’ and ‘content length’. Some of them are unstable, such as ‘total votes’ and ‘reviewer ranking’. The ERO features are selected by comparing the average ERO value with empirical thresholds, which is 0.01 and -0.01 in this work.

3.6 Experiment and Results

The proposed ERO analysis is implemented and experiments are conducted in this section. We first build a web crawling system and crawl review data from Amazon. ERO analysis is applied and suspicious products are picked up. Then, expert reviews are employed to evaluate the detection performance. The ERO analysis is implemented in Matlab, and the crawling system is build in Python.

3.6.1 Performance Evaluation

Unlike fraudulence detection in a traditional system, such as email spam, it is often hard to get ground truth for fake reviews in a product/service review system. Lack of ground truth is one of challenges when study online review fraudulence. To our best knowledge, there is no research work that can perfectly address this issue. However, researchers have made several attempts to evaluate the performance of a fake review detection.

Previous research employed different approaches to obtain ground truth data. There are mainly three types of approaches. *First*, some early work manually inspect reviews and extract simple features such as duplicated and near-duplicated

Table 10: Review feature crawled from Amazon

Product	Review	Reviewer
ASIN	Rating value	Customer ID
Average rating	# of helpful votes	Reviewer ranking
Category	# of total votes	# of helpful received votes
	Review date	# of total received votes
	Review text	# of reviews
	Verified review or not	# of verified reviews

reviews or unexpected rating patterns [19]. This type of approach is limited since it largely depends on heuristics and the assumption that fake review structures are known. *Second*, a few researchers create ground truth data by hiring people to write fake reviews [25, 28]. They then develop detectors that compare the features of genuine and fake reviews. Although these classifiers have performed well with those artificial data, it is questionable whether datasets generated by hired people can be representative of actual fake reviews in practice. The *third* type is employing expert reviews. For example, Mukherjee et al. generated ground truth by hiring experts to manually detect fake reviews given some intuitive features and hints [27]. In this work, we adopt the third approach, in which we employ expert reviewers to evaluate the detection accuracy.

In addition, we argue that the detection accuracy should be presented as **False Alarm Rate** (FAR). Although the other metric – Detection Rate (DR) might be more intuitive, it is hard to estimate due to the lack of ground truth. Fake reviews are complex and multidimensional. It is extremely difficult for a single method to achieve high accuracy. Multiple detection methods have to be used together to capture different aspects of fake reviews.

3.6.2 Dataset

We develop web crawlers to collect real data from Amazon. The crawlers are implemented in Python [29]. The dataset we use in the experiments are collected

Table 11: Summary of dataset

Number of products	916
Number of reviews	418549
Average product rating	3.98
Product categories	‘Toys and Games’ and ‘Electronics’

under several guidelines.

1. The number of reviews is greater than 50.
2. The average rating is between 2.5 and 4.8.
3. Product category is either ‘Toys and Games’ or ‘Electronics’.
4. We only collect the first 40 pages of each category.

For each product, there are 3 types of information crawled.

1. Product meta-information. The product ASIN (Amazon Standard Identification Number), average rating and category are obtained.
2. Review information, such as rating value, date, text, number of votes, etc. are crawled from Amazon.
3. Reviewer information, such as customer ID, reviewer ranking, number of reviews, etc. are crawled from Amazon.

Table 10 shows the detail information we obtained. There are 916 products collected. A summary of the dataset is shown in Table 11.

3.6.3 Victim Product Detection Comparison

The ‘verified’ feature and ‘day of week’ feature are used in the experiment. In this section, we compare the ERO analysis results with two common approaches. One is the consistency detection that is described in Section 3.4.1. The other is duplication detection described in [19].

Table 12: Experiment parameters

Approach	Parameters
ERO	$w = 2$
Consistency	$\mu_i = \text{average rating}$ $\nu = 0.6$ $\bar{h}_0 = 3$

All the 916 products are analyzed by the three detection methods. For consistency detection, the trigger threshold we used is $\nu = 0.6$, and the minimum threshold $\bar{h}_0 = 3$. More details of consistency detection can be found in Section 3.4.1. In ERO analysis, the *boxplot* method is applied on the 916 products, by which the detection thresholds are $\theta_{upper} = 0.16$, $\theta_{lower} = -0.17$ for ‘day of week’ feature, and $\theta_{upper} = 0.1$, $\theta_{lower} = -0.15$ for ‘verified’ feature. There are 16 products detected to be suspicious. We compare the pdetection results, and display the comparison in Fig 14. The ERO analysis detects 16 suspicious products, including 7 products detected by ‘day of week’ feature and 9 products detected by ‘verified’ feature. The consistency method detects 28 products, while the duplication method detects 123 products. We have following observations.

- There are small overlaps among the three methods. One product is detected by both ERO analysis and duplication method, and two products are detected by both consistency and duplication methods. Fake reviews have very diverse structures. Single method is not able to detect all those structures. ERO analysis is important if the false alarm rate is relatively low, as it provides a new angle that missed by existing methods, such as duplication method and consistency method.
- There are 123 products detected by duplication method. We look at the products detected by duplication method and find that some duplicated texts are just “5 star”, “excellent”, “I like it” and some common phrases. Without

Figure 14: Fake review detection comparison

Table 13: Expert review

	Verified	Day of Week	Total
False Alarm Rate	11.1%	0	6.3%
Confidence Change	8.6%	12.3%	10.5%

further information, it is hard to tell whether they are fake. It is highly possible that the duplication method has a high false alarm rate. There are 16 products detected by the ERO analysis. The detection rate may be low. However, considering the diverse of fake review structures and the specific angle captured by ERO analysis, the low detection rate is reasonable as long as the false alarm rate is relatively high.

It is noticed that the thresholds of those detection methods impact the detection rate and overlaps. In this experiment, the consistency method uses the thresholds proposed in [25]. In practice, detection thresholds should be well tuned when multiple methods are used to capture the different aspects of fake reviews.

3.6.4 Expert Reviews

In the experiment of previous subsection, there are 16 products detected by ERO analysis. As we discussed in Section 3.6.1, no ground truth is available for performance evaluation. Therefore, in order to evaluate the results of ERO

analysis, we conduct expert reviews in this section.

The detection results are presented as one or two sentences and provided to the expert reviewers as hints. The expert reviewers answer two questions – “*do you think the product contains fake reviews?*” and “*how confident do you feel when you answer the first question?*”. There are 16 expert reviewers participating in the evaluation. When determine the results of expert review, we use majority votes. Particularly, if 8 or more people vote the product as non-victim, then the detection is considered as false alarm.

The results are shown in Table 13. According to the expert reviews, 6.3% detected products are voted as false alarm. Particularly, 11.1% products detected by ‘verified’ feature are voted as false alarm, and no products detected by ‘day of week’ feature are voted as false alarm. It is noticed that, ‘day of week’ feature is more convincing than ‘verified’ feature. We also investigate the confidence change. The products are first provided to expert reviewers without any hints. After they making their decisions, the hints are provided to them and they make their decision again. For each decision, they choose one confidence level from the 5 levels. We calculate the confidence change after they getting the hints, and the results are shown in Table 13. There is an average 10.5% confidence increase, which means people can feel more confident if they can consult the ERO analysis.

3.6.5 Discussion

When a smart attacker learns ERO principle and ERO analysis, the attacker may be able to perform certain actions to avoid the detection of ERO analysis. For example, there are some fake reviewers working on specific day, and this pattern can be detected by ERO analysis with ‘day of week’ feature. However, a smart attacker can change the attack strategy to avoid the detection of ERO analysis, such as posting fake reviews randomly in a week. In this case, the ERO analysis

is not attack resistance.

To our best knowledge, almost all the existing approaches cannot guarantee attack resistance. However, the proposed ERO analysis can increase the cost of attack and thus reduce the impact of fake reviews. For example, if the reviewers' IP addresses are available for ERO analysis, then we are able to estimate the reviewers' geographic location information based on their IP addresses. The ERO analysis is able to detect correlations between rating value and reviewers' city or state. Although it is still not attack resistance, as attackers can technically change their IP addresses, the attack cost must be increased.

In practice, a more effective way to improve attack resistance is using multiple detection methods to capture the diverse structures of fake reviews.

3.7 Conclusion

The ERO principle, which is a new angle to detect fake reviews, has been introduced in this manuscript. The fraudulence signal detected by ERO analysis is very different from the traditional approaches. Furthermore, it needs very limited data to set up detection thresholds, and only requires the reviews for a particular product in order to determine whether this product is under review manipulation. Therefore, it can be performed in real-time. Experiments based on real data are conducted. Several review features are examined for the applicability of ERO analysis, and two of them are found applicable. The detection results are compared with two common existing methods. The proposed ERO analysis provides a new angle of fake review detection. We also conduct expert reviews to evaluate the ERO analysis results, in which the ERO analysis gives a low false alarm rate. In the future, we look forward to selecting more ERO features and implementing a real application and involve more users for testing.

List of References

- [1] Q. Fottrell, “Yelp deems 20% of user reviews ‘suspicious’,” 2013, visited on 2015-08-09. [Online]. Available: http://www.marketwatch.com/story/20-of-yelp-reviews-are-fake-2013-09-24?link=MW_story_latest_news
- [2] S. Bennett, “The rise of social media in e-commerce,” 2014, visited on 2015-11-18. [Online]. Available: <http://www.adweek.com/socialtimes/social-e-commerce/496484>
- [3] “Final pre-christmas push propels u.s. online holiday season spending through december 26 to record \$30.8 billion,” 2010, visited on 2015-01-20. [Online]. Available: <http://ir.comscore.com/releasedetail.cfm?ReleaseID=539354>
- [4] J. Morgan and J. Brown, “Reputation in online auctions: The market for trust,” *California Management Review*, vol. 49, no. 1, pp. 61–81, 2006.
- [5] A. HARMON, “Amazon glitch unmasks war of reviewers,” 2004, visited on 2015-01-20. [Online]. Available: <http://www.nytimes.com/2004/02/14/us/amazon-glitch-unmasks-war-of-reviewers.html>
- [6] M. Abadi, M. Burrows, B. Lampson, and G. Plotkin, “A calculus for access control in distributed systems,” *ACM Trans. Program. Lang. Syst.*, vol. 15, no. 4, pp. 706–734, 1993.
- [7] A. Josang, R. Ismail, and C. Boyd, “A survey of trust and reputation systems for online service provision,” *Decision Support Systems*, vol. 43, no. 2, pp. 618 – 644, 2007.
- [8] P. Laureti, L. Moret, Y.-C. Zhang, and Y.-K. Yu, “Information filtering via iterative refinement,” *EPL (Europhysics Letters)*, vol. 75, no. 6, p. 1006, 2006.
- [9] B. Yu and M. P. Singh, “An evidential model of distributed reputation management,” in *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1*, ser. AAMAS ’02. ACM, 2002, pp. 294–301.
- [10] Y. Liu and Y. Sun, “Anomaly detection in feedback-based reputation systems through temporal and correlation analysis,” in *2010 IEEE Second International Conference on Social Computing (SocialCom)*, Aug 2010, pp. 65–72.
- [11] C. Dellarocas, “Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior,” in *Proceedings of the 2Nd ACM Conference on Electronic Commerce*, ser. EC ’00. ACM, 2000, pp. 150–157.
- [12] N. Jindal and B. Liu, “Opinion spam and analysis,” in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ser. WSDM ’08. ACM, 2008, pp. 219–230.

- [13] Yelp, “Yelp’s review filter explained,” 2010, visited on 2015-03-21. [Online]. Available: <http://officialblog.yelp.com/2010/03/yelp-review-filter-explained.html>
- [14] Yelp. “Why yelp has a review filter.” Visited on 2015-03-21. 2009. [Online]. Available: <http://officialblog.yelp.com/2009/10/why-yelp-has-a-review-filter.html>
- [15] A. Josang and R. Ismail, “The beta reputation system,” in *Proceedings of the 15th bled electronic commerce conference*, 2002, pp. 41–55.
- [16] W. Jianshu, M. Chunyan, and G. Angela, “An entropy-based approach to protecting rating systems from unfair testimonies,” *IEICE TRANSACTIONS on Information and Systems*, vol. 89, no. 9, pp. 2502–2511, 2006.
- [17] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, “Finding deceptive opinion spam by any stretch of the imagination,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT ’11, 2011, pp. 309–319.
- [18] R. Y. K. Lau, S. Y. Liao, R. C.-W. Kwok, K. Xu, Y. Xia, and Y. Li, “Text mining and probabilistic language modeling for online review spam detection,” *ACM Trans. Manage. Inf. Syst.*, vol. 2, no. 4, pp. 25:1–25:30, 2012.
- [19] E. Choo, T. Yu, M. Chi, and Y. Sun, “Revealing and incorporating implicit communities to improve recommender systems,” in *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, ser. EC ’14. ACM, 2014, pp. 489–506.
- [20] G. Wang, S. Xie, B. Liu, and P. Yu, “Review graph based online store review spammer detection,” in *2011 IEEE 11th International Conference on Data Mining (ICDM)*, Dec 2011, pp. 1242–1247.
- [21] A. Mukherjee, B. Liu, J. Wang, N. Glance, and N. Jindal, “Detecting group review spam,” in *Proceedings of the 20th International Conference Companion on World Wide Web*, ser. WWW ’11. ACM, 2011, pp. 93–94.
- [22] Y. Yang, Y. Sun, S. Kay, and Q. Yang, “Securing rating aggregation systems using statistical detectors and trust,” *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 4, pp. 883–898, 2009.
- [23] “Reviewpro,” visited on 2014-11-19. [Online]. Available: <http://www.reviewpro.com/>
- [24] “Trustyou,” visited on 2015-01-10. [Online]. Available: <http://www.trustyou.com/>

- [25] Y. Liu, Y. Sun, and T. Yu, “Defending multiple-user-multiple-target attacks in online reputation systems,” in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, Oct 2011, pp. 425–434.
- [26] M. Frigge, D. C. Hoaglin, and B. Iglewicz, “Some implementations of the boxplot,” *The American Statistician*, vol. 43, no. 1, pp. 50–54, 1989.
- [27] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh, “Spotting opinion spammers using behavioral footprints,” in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’13. ACM, 2013, pp. 632–640.
- [28] M. Ott, C. Cardie, and J. T. Hancock, “Negative deceptive opinion spam.” in *HLT-NAACL*, 2013, pp. 497–501.
- [29] Y. Zeng, Y. Zhu, and Y. Sun, “Reviewsec: Security analysis tool for online reviews,” *STCSN E-Letter*, vol. 2, no. 4, 2014.