A SURVEY ON DETECTION AND PREVENTION OF CELLULAR APP RANKING SCAMS IN AN APP STORE

Sumaiya.s[1], Mrs.Anitha.T.N[2]

[1]*MTech 4th Sem Student, Department of Computer Science and Engineering, SJCIT, Karnataka, India*

[2]*Asso.Professor, Department of Computer Science and Engineering, SJCIT, Karnataka, India*

ABSTRACT

In the recent years, Cellular Apps are become popular and growing at an amazing rate due to advancement in Cellular technology and Cellular devices. Cellular App ranking Scams refers to fraudulent or vulnerable activities which have a purpose of bumping up the Apps in the fame or leader board list. Many shady means are used more frequently by App developers, such as expanding their Apps' business or posting imposter App evaluations, to confer positioning misrepresentation. While the importance and necessity of preventing ranking Scams has been widely recognized, there is limited understanding in this area. To this end, the paper proposes a detection of Cellular App ranking Scams in App store. The proposed system mines the active periods such as leading sessions of Cellular Apps to accurately locate the ranking Scams. These leading sessions can be useful for detecting the local anomaly instead of global anomaly of App rankings. Besides this, by modeling Apps ranking, rating and review behaviours using statistical hypotheses tests, Investigate three types of evidences, they are ranking based evidences, rating based evidences and review based evidences. Furthermore, propose an aggregation method based on optimization to integrate all the evidences for Scam detection. Finally, the proposed system will be evaluated with real-world App data which is to be collected from the App Store for a long time period.

Keywords*: - Cellular Apps, ranking Scam detection, evidence aggregation, historical ranking records, rating and review.*

I. INTRODUCTION

The quantity of Cellular Apps has developed at a stunning rate in the course of recent years. Such as, there are more than 1.6 million Apps at Apple's App store and Google Play at the end of April 2013.To fortify the advancement of Cellular Apps, numerous App stores launched day by day App leader boards, which show the graph rankings of most well-known Apps. In reality, the App pioneer board is a standout amongst the most vital routes for advancing Cellular Apps. A top rank on the pioneer board usually leads to a huge number of downloads and million dollars in revenue. As a result, many App developers incline to explore various ways such as advertisement drive to promote their Apps to get higher position in such App leader boards. Rather than depending on conventional advertising arrangements, shady App designers resort to some fake intends to purposely help their Apps and in the long run control the outline rankings on an App store. This is typically executed by utilizing supposed "bot farmstead" or "human water armed forces" to blow up the App downloads, appraisals and audits in a brief while. Leading events of Cellular Apps forms different leading sessions. The Cellular Apps not always ranked high in the leader boards, but it usually happens in the leading sessions. So, detecting ranking Scam of Cellular Apps is actually the process to detect it within the leading session of the Cellular Apps. Especially, this paper proposes a simple and effective algorithm to recognize the leading sessions of each Cellular App based on its historical ranking records. This is one of the Scam evidence. Also, two types of Scam evidences are proposed based on Apps' rating and review history, which gives some anomaly patterns from Apps' historical rating and review records. In addition, we propose an unsupervised evidence aggregation method to consolidate these three types of evidences for assessing the credibility.

2. LITERATURE SURVEY

The research work of this study is divided into three categories. They are i) web ranking spam detection [1], [2], [3], ii) online review spam detection [4], [5], [6] and iii) mobile App recommendation [7], [8], [9]. The first category is Web ranking spam detection. The web ranking spam refers to any intentional actions which bring to selected web pages an inexcusable auspicious relevant importance. Following is the work done on web ranking spam detection. Zhou *et al* [1] have studied the problem of unsupervised Web ranking spam detection. Specifically, they proposed an efficient online link spam and term spam detection methods using spamicity. Ntoulas *et al.* [2] have studied various aspects of content-based spam on the Web and presented a number of heuristic methods for detecting content based spam. Recently, Spirin *et al.* [3] have reported a survey on Web spam detection, which comprehensively introduces the principles and algorithms in the literature. Indeed, the work of Web ranking spam detection is mainly based on the analysis of ranking principles of search engines, such as Page Rank and query term frequency. This is different from ranking fraud detection for mobile Apps. They categorize all existing algorithms into three categories based on the type of information they use: content-based methods, link-based methods, and methods based on non-traditional data such as user behaviour, clicks, and HTTP sessions. In turn, there is a sub categorization of link-based category into five groups based on ideas and principles used: labels propagation, link pruning and reweighting, labels refinement, graph regularization, and feature based.

The second category is focused on detecting online review spam. Lim *et al.* [4] have identified several representative behaviors of review spammers and model these behaviors to detect the spammers. This paper aims to detect users generating spam reviews or review spammers. They identify several characteristic behaviors of review spammers and model these behaviors so as to detect the spammers. In particular, authors seek to model the following behaviors. First, spammers may target specific products or product groups in order to maximize their impact. Second, they tend to deviate from the other reviewers in their ratings of products. They propose scoring methods to measure the degree of spam for each reviewer and apply them on an Amazon review dataset.

Authors then select a Subset of highly suspicious reviewers for further scrutiny by user evaluators with the help of a web based spammer evaluation software specially developed for user evaluation experiments. Wu *et al.* [5] have studied the problem of detecting hybrid shilling attacks on rating data. The proposed approach is based on the semi-supervised learning and can be used for trustworthy product recommendation. This paper presents a Hybrid Shilling Attack Detector or HySAD for short, to tackle these problems. In particular, HySAD introduces MC-Relief to select effective detection metrics, and Semi- supervised Naive Bayes (SNBλ) to precisely separate Random-Filler model attackers and Average-Filler model attackers from normal users. Xie *et al.* [6] have studied the problem of singleton review spam detection. Specifically, they solved this problem by detecting the co-anomaly patterns in multiple review based time series. Although some of above approaches can be used for anomaly detection from historical rating and review records, they are not able to extract fraud evidences for a given time period (i.e., leading session).

Finally, the third category includes the studies on mobile App recommendation. Yan *et al.* [7] developed a mobile App recommender system, named Appjoy, which is based on user's App usage records to build a preference matrix instead of using explicit user ratings. Also, to solve the sparsity problem of App usage records. Shi *et al.* [8] studied several recommendation models and proposed content based collaborative filtering model, named Eigenapp, for recommending Apps in their Web site Getjar. In addition, some researchers studied the problem of exploiting enriched contextual information for mobile App recommendation. Zhu *et al.* [9] proposed a uniform framework for personalized context-aware recommendation, which can integrate both context independency and dependency assumptions. However, to the best of our knowledge, none of previous works has studied the problem of ranking Scam detection for Cellular Apps. Detection of ranking Scam for Cellular Apps is still under a subject to research. To fill this crucial lack, we propose to develop a ranking Scam detection system for Cellular Apps. We also determine several important challenges. First challenge, in the whole life cycle of an App, the ranking Scam does not always happen, so we need to detect the time when Scam happens. This challenge can be considered as detecting the local anomaly in place of global anomaly of Cellular Apps. Second challenge, it is important to have a scalable way to positively detect ranking Scam without using any basis information, as there are huge number of Cellular Apps, it is very difficult to manually label ranking Scam for each App. Finally, due to the dynamic nature of chart rankings, it is difficult to find and verify the evidences associated with ranking Scam, which motivates us to discover some implicit Scam patterns of Cellular Apps as evidences.

3. PROPOSED SYSTEM

With the increase in the number of web Apps, to detect the fraudulent Apps, we have propose a simple and effective algorithm which identifies the leading sessions of each App based on its historical ranking of records. By analysing the ranking behaviours of Apps, we discover that the fraudulent Apps often have different ranking patterns in each leading session compared with normal Apps. Thus, we identify some Scam evidences from Apps' historical ranking records and develop three functions to obtain such ranking based Scam evidences.

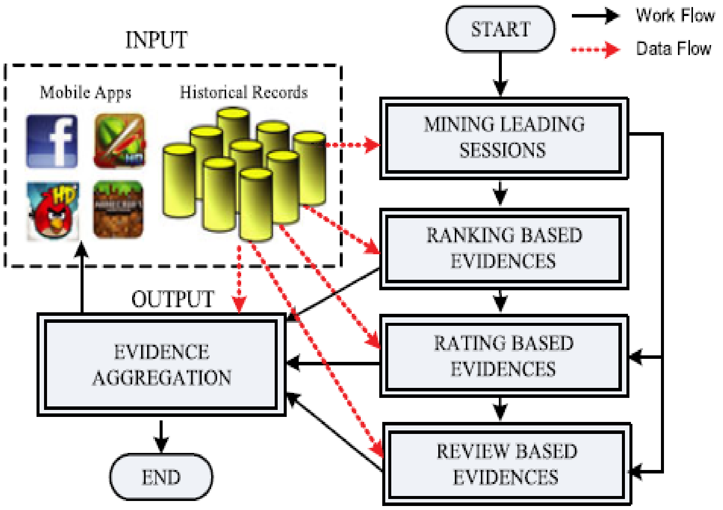

Fig -1*:* Ranking Scam detection system framework

Further, we propose two types of Scam evidences based on Apps' rating and review history. It reflects some anomaly patterns from Apps' historical rating and review records. The leading sessions of Cellular App signify the period of popularity, and so these leading sessions will comprise of ranking manipulation only. Hence, the issue of identifying ranking Scam is to identify vulnerable leading sessions. Along with this, the main task is to extract the leading sessions of a Cellular App from its historical ranking records.

Module 1: Leading Events

Given a positioning limit K _ 2 [1, K] a main occasion e of App a contains a period range also, relating rankings of a, Note that positioning edge K * is applied which is normally littler than K here on the grounds that K may be huge (e.g., more than 1,000), and the positioning records past K _(e.g., 300) are not exceptionally helpful for recognizing the positioning controls. Moreover, it is finding that a few Apps have a few nearby driving even which are near one another and structure a main session.

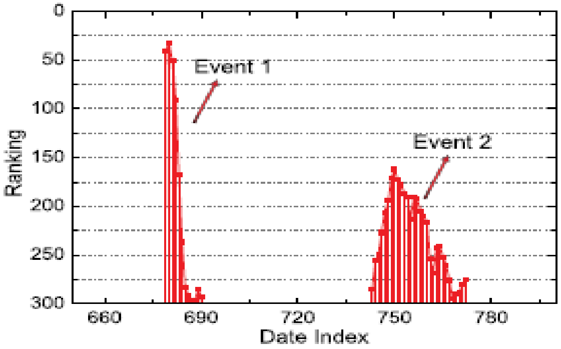

Fig -2: Example of Leading Events

Module 2: Leading Sessions

Instinctively, mainly the leading sessions of Cellular app signify the period of popularity, and so these leading sessions will comprise of ranking manipulation only. Hence, the issue of identifying ranking Scam is to identify deceptive leading sessions. Along with the main task is to extract the leading sessions of a Cellular App from its historical ranking records.

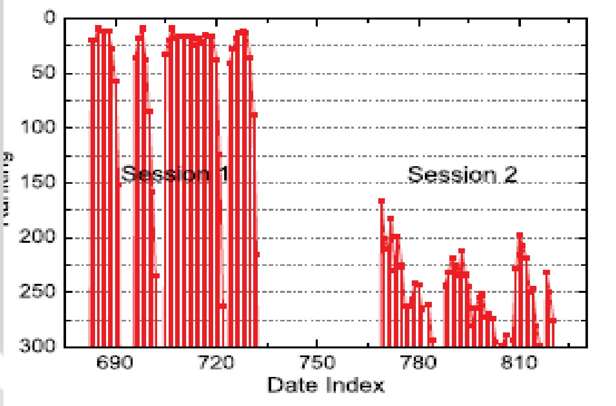

Fig -3: Example of Leading Sessions

Module 3: Identifying the Leading Sessions for Cellular apps

Basically, mining leading sessions has two types of steps concerning with Cellular Scam apps. Firstly, from the Apps historical ranking records, discovery of leading events is done and then secondly merging of adjacent leading events is done which appeared for constructing leading sessions. Certainly, some specific algorithm is demonstrated from the pseudo code of mining sessions of given Cellular App and that algorithm is able to identify the certain leading events and sessions by scanning historical records one by one.

Module 4: Identifying Evidences for Ranking Scam detection

4.1 Ranking Based Evidence

It concludes that leading session comprises of various leading events. Hence by analysis of basic behaviour of leading events for finding Scam evidences and also for the app historical ranking records, it is been observed that a specific ranking pattern is always satisfied by app ranking behaviour in a leading event.

4.2 Rating Based Evidence

Resolving the problem of "restrict time reduction", identification of Scam evidences is planned due to app historical rating records. As we know that rating is been done after downloading it by the user, and if the rating is high in leader board considerably that is attracted by most of the Cellular app users. Spontaneously, the ratings during the leading session gives rise to the anomaly pattern which happens during rating Scam. These historical records can be used for developing rating based evidences.

4.3 Review Based Evidence

Review which contains some textual comments as reviews by app user and before downloading or using the app user mostly prefer to refer the reviews given by most of the users. Therefore, although due to some previous works on review spam detection, there still issue on locating the local anomaly of reviews in leading sessions. So based on apps review behaviors, Scam evidences are used to detect the ranking Scam in Cellular app.

4. CONCLUSION

This paper reviews various existing methods used for web spam detection, which is related to the ranking Scam for Cellular Apps. Also, we have seen references for online review spam detection and Cellular App recommendation. By mining the leading sessions of Cellular Apps, we aim to locate the ranking Scam. The leading sessions works for detecting the local anomaly of App rankings. The system aims to detect the ranking Scam based on three types of evidences, such as ranking based evidences, rating based evidences and review based evidences. Further, an optimization based aggregation method combines all the three evidences to detect the Scam.

5. REFERENCES

[1].B. Zhou, J. Pei, and Z. Tang. A spamicity approach to web spam detection. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, SDM'08, pages 277–288, 2008.

[2].A. Ntoulas, M. Major, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 83–92, 2006.

[3]. N. Spirin and J. Han. Survey on web spam detection: principles and algorithms. *SIGKDD Explor. Newsl*, 13(2):50–64, May 2012.

[4].E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 939–948, 2010.

[5]. Z.Wu, J.Wu, J. Cao, and D. Tao. Hysad: a semi- supervised hybrid shilling attack detector for trustworthy product recommendation. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 985–993, 2012.

[6]. S. Xie, G. Wang, S. Lin, and P. S. Yu. Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 823–831, 2012.

[7].B. Yan and G. Chen. Appjoy: personalized mobile application discovery. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*, MobiSys '11, pages 113– 126, 2011.

[8]. K. Shi and K. Ali. Getjar mobile application recommendations with very sparse datasets, In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* KDD '12, pages 204–212, 2012.

[9]. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in VLDB, 1994.

[10]. H. Zhu, E. Chen, K. Yu, H. Cao, H. Xiong, and J. Tian. Mining personal context-aware preferences for mobile users. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1212–1217, 2012.

[11].Hengshu Zhu, Hui XiongDiscovery of Ranking Fraud for mobile users. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1212–1217, 2012.

[12] .D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., pp. 993–1022, 2003.

[13] .Y. Ge, H. Xiong, C. Liu, and Z.-H. Zhou, "A taxi driving fraud detection system," in Proc. IEEE 11th Int. Conf. Data Mining, 2011, pp. 181–190.

[14]. D. F. Gleich and L.-h. Lim, "Rank aggregation via nuclear norm minimization," in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2011, pp. 60–68.

[15]. T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proc. Nat. Acad. Sci. USA, vol. 101, pp. 5228–5235, 2004.

[16]. G. Heinrich, Parameter estimation for text analysis, "Univ. Leipzig,Leipzig,Germany,Tech.Rep., http://faculty.cs.byu.edu/~ringger/CS601R/papers/Heinrich-GibbsLDA.pdf, 2008.

[17]. N. Jindal and B. Liu, "Opinion spam and analysis," in Proc. Int.Conf. Web Search Data Mining, 2008, pp. 219–230.

[18]. A. Klementiev, D. Roth, and K. Small, "An unsupervised learning algorithm for rank aggregation," in Proc. 18th Eur. Conf. Mach. Learn., 2007,pp. 616–623.

[19] A. Klementiev, D. Roth, and K. Small, "Unsupervised rank aggregation with distance-based models," in Proc. 25th Int. Conf. Mach. Learn.,2008, pp. 472–479.

[20] A. Klementiev, D. Roth, K. Small, and I. Titov, "Unsupervised rank aggregation with domain-specific expertise," in Proc. 21st Int. Joint Conf. Artif. Intell., 2009, pp. 1101–1106 .

[21] http://en.wikipedia.org/wiki/cohen's kappa.

[22] https://developer.apple.com/news/index.php?id=0-2062012a.

[23]http://venturebeat.com/2012/07/03/apples crackdown-on-app-ranking-manipulation/.

[24]http://www.ibtimes.com/apple-threatens-crackdown-biggest-app-store-ranking-fraud-406764.

[25] http://www.ibtimes.com/apple-threatens- crack down- biggestapp-store-ranking-fraud-406764.

[26] L. Azzopardi, M. Girolami, and K. V. Risjbergen Investigating the relationship between language model perplexity and ir precision recall measures. . In Proceedings of the 26th International Conference on Research and Development in Information Retrieval (SIGIR'03), pages 369–370, 2003