Compromised Account Detection Based on Clickstream Data

Tobias Weller*
Institute AIFB (KIT)
Karlsruhe, Germany
tobias.weller@kit.edu

ABSTRACT

The number of users of the world wide web is constantly increasing. However, this also increases the risks. There is the possibility that other users illegally gain access to a users' account of social networks, web shops or other web services. Previous work use graph-based methods to identify hijacked or compromised accounts. Most often posts are used in social networks to detect fraudulences. However, not every compromised account is used to spread propaganda information or phishing attacks. Therefore, we restrict ourselves to the clickstreams from the accounts. In order to identify compromised accounts by means of clickstreams, we will also consider a temporal aspect, since the preferences of a user change over time. We choose a hybrid approach consisting of methods from subsymbolic and symbolic AI to detect fraudulences in clickstreams. We will also take into account the experience of domain experts. Our approach can also be used to identify not only compromised accounts but also shared accounts on instance streaming sites.

CCS CONCEPTS

• **Information systems** → **Web log analysis**; • **Security and privacy** → **Intrusion/anomaly detection and malware mitigation**; • **Social and professional topics** → *Identity theft*; • **Mathematics of computing** → Probabilistic representations;

KEYWORDS

Clickstream Analysis, Clickstream Fraud Detection, Anomaly Detection, Machine Learning

ACM Reference Format:
Tobias Weller. 2018. Compromised Account Detection Based on Clickstream Data. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3184558.3186569

1 PROBLEM

The Internet serves as a worldwide interconnection of individual networks. The number of users participating in this network increased since its beginning in 1994. Currently, 4,157 million users are estimated to use the world wide web[1]. At the same time are the number of users in social networks[2] such as Facebook and Twitter increasing, as is the number of digital buyers[3] in web shops such as Amazon and Alibaba. All these networks have in common that users have to login to use the application in full. These accounts are in particular interested for hackers. Hackers try to gain unnoticed access to the users' accounts in order to use them for their criminal activities. These hijacked accounts are among others used for phishing attacks, cyber crime-related scams, spam campaigns and spreading propaganda information.

However, there is also a preliminary stage, even before hijacked accounts, namely compromised accounts. Compromised accounts are those accounts whose passwords are unnoticedly available to others, so that these people can gain access to the account unnoticed. In this case, the hackers often just try to gather as much information as possible and log out, without posting or to inflict further harm. Often the affected users are not aware of the fact that their account has been compromised. Therefore are indicators and measures needed to identify compromised accounts or abnormally movements of users. Methods that evaluate the users' postings to determine a compromise are unsuitable for this use case. Current approaches like sending emails when logging in from unknown clients, like Facebook or Google does, are only of limited use. Users are annoyed by those emails or remain unnoticed in the inbox.

A related use case, which is similar, is the detection of shared accounts. Often are users sharing their accounts so that two or more persons use the same account. On the one hand, provider of web solutions may not be interested in users sharing their accounts with others, but on the other hand, it might be interesting for those providers to identify the currently browsing person and make targeted advertising or recommendations. This way, you could advertise or recommend on the basis of the person who is currently using the account and not on the basis of the account itself. For example, when using a family Amazon account you could try to identify and address the person yourself. Figure 1 summarizes the two possible use cases.

2 STATE OF THE ART

Within this work, three topics are dealt with. These are fraud and anomaly detection, web log analysis and to a small extent data modeling and reasoning. In the following we will show how current research is dealing with these topics.

In the area of socially compromised accounts, the posts of the users are often analyzed to detect changes of content, as a compromise indication [8]. Most work investigates behavioral changes. Thereby statistical methods are used to recognize these. Bayesian

*Supervised by Dr. Maria Maleshkova
[1]http://www.internetworldstats.com/emarketing.htm, last access: 23rd February 2018

[2]https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/, last access: 23rd February 2018
[3]https://www.statista.com/statistics/251666/number-of-digital-buyers-worldwide/, last access: 23rd February 2018

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186569

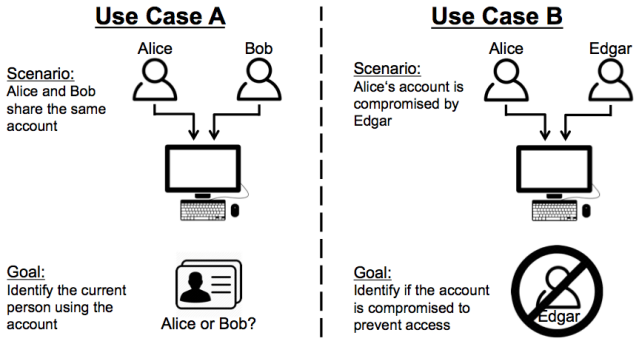

Figure 1: Two use cases can be tackled with this work. Use Case A: Identification of a shared account person. Use Case B: Detection of a compromised account.

models are used to identify anomalies in graphs at discrete times [11]. Hereby, the network structure of the social network is exploited. Similarity analysis is also used to detect compromises in individual accounts [7]. In contrast to fraud detection in social networks, similar methods are used to detect anomalies in online advertising. Providers of an Internet platform receive a commission for each successful click on an advertisement. The providers try to increase their commission by fraudulent clicks in these so-called pay-per-click models. Subject of the research was to identify these fraudulent clicks [27]. Here, however, the focus is on identifying duplicates rather than a sequence of clicks. Association rules are also used for fraud detection in advertising networks [13]. Just as in social networks, graph-based methods are used to detect anomalies in graph-based data. A few works often use two techniques [17]. On the one hand a technique to identify regularities in graphs to identify the normality. And secondly, techniques to identify anomalies, i. e. deviations from the standard.

Fraud detection is also subject in identifying anomalies in UNIX commands. Hereby, sequences of commands of users are compared by using similarity measures to the profile of an user's command sequence [12]. Other techniques on the server side use text mining methods for intrusion detection [2]. Another technique is the use of Neural Networks for intrusion detection. This is called NNID (Neural Network Intrusion Detector) [22]. Here, the presumption is that each user leaves a fingerprint on the server, which is basically the same assumption as we do, but ours is based on clickstreams. A Neural Network is used to learn the print and identify users, based on these prints. If a user's behavior does not match to his print, then his account is classified as a possible security breech. Surveys provide a comprehensive overview of anomaly and intrusion detection systems. In addition, trends are discussed [21, 24].

Besides the identification of anomalies of a user in social networks or on servers, fraud detection was often used in the financial sector. Many works use neural networks for fraud detection. Some of them operate in online systems [6]. Besides this, data mining techniques and neural network algorithms were combined successfully to obtain a high fraud coverage, combined with a low false alarm rate [4]. Within these works, the time aspect in financial fraud detection is always considered [10]. Surveys in the field of financial fraud detection summarizes applied methods from the past [5, 26]. The surveys shows that mining algorithms, statistical tests, regression analysis, neural networks, decision trees and Bayesian networks are used for fraud detection. Moreover, the surveys show that in general, the detecting effect and accuracy of neural networks are superior to regression model.

As seen, clickstreams to identify compromised accounts has not been subject matter of research. However, web logs have been analyzed for various reasons. AltaVista Search Engine query logs had been analyzed. Hereby, correlation analysis were used to analyze log entries and studying the interaction of terms within queries [25]. A further analysis of transaction logs of a korean web search engine (NAVER) shows that user's behave in a simple way [20]. Other work uses unsupervised algorithms to cluster users, based on online transaction data from an university. In addition, filters and combinations had been presented [9]. Matrix clustering was successfully implemented for representing relationships between pages and users in a binary matrix from Web access logs. The page clusters extracted by matrix clustering can be applied to web access prediction [19]. An overview of about 10 years of research on log analysis had been presented in surveys [3].

To a small extent, we will use methods of logic and modeling. Organizations analyze such data to evaluate the effectiveness of their campaigns and applications [14]. Further papers focuses to use the web data for fuzzy approximate reasoning for recommendation systems [16]. Dynamic multinomial probit model of clickstreams are used as model for predicting and categorizing clickstream paths. It has been shown that this technique outperforms traditional first-order Markov models [15]. Besides considering only clickstreams, further features are taken into account like e.g. recommendation lists, ratings, styles and tags. Considering additional features shows a significant impact, however both, positive and negative [18]. Other work identified typical and atypical sessions in clickstreams. The outliers can be identified with different distance measures such as the Mahalanobis distance in the user session space. The results demonstrate that identifying typical and atypical user sessions is extremely valuable for cleaning "noisy" user session data for increased accuracy in evaluating user experience [23]. Semantic user models are also the subject of research. Words can be extracted from the user's session and disambiguate with words from Wordnet or other lexicons. By means of similarity measurements, the semantic vector spaces can then be classified [1].

3 PROPOSED APPROACH

The overall goal of this work is to detect compromised user accounts based on clickstream data. Currently, there is very few research made in the area of detecting compromised user accounts. Mostly, the work focuses on detecting hijacked social accounts or the intrusion on web servers. Moreover, the amount of work considering clickstream data to detect compromised accounts is even less. Existing approaches mostly uses graph-based methods to identify hijacked accounts. However, we assume that the clickstreams of a compromised account differ from the clickstreams made before the compromising of the account. The hacker, which has gained access to the account, traverses and behaves differently than the actual user. These produced clickstreams therefore differ from the

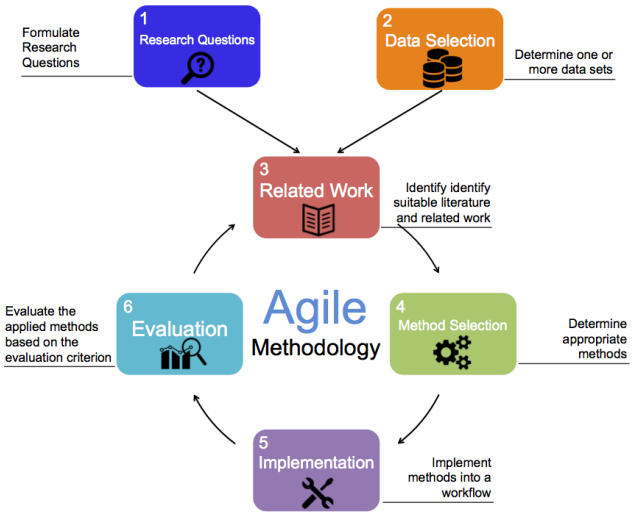

Figure 2: Agile approach for the underlying work.

ones before and can be considered as anomalies. Besides detecting compromised accounts, we will detect if the account is a shared account and if so, the current person using the account. Knowing the person, using the account, allows for personalized advertisements, even in a shared account manner. For this purpose we will also use clickstream data. However, we also have to consider the user's preferences over time, as they can change. Only in this way can we identify unforeseen things and determine whether the preferences are too different. Accounts whose preferences change abruptly or do not match are marked as conspicuous.

Throughout the entire work, we use an agile approach to obtain initial results as quickly as possible in order to gain initial insights and use this to adapt the methods and improve the models. We hope that this approach will produce good results as soon as possible and reject erroneous hypotheses as soon as possible. Figure 2 shows an high level overview of the agile process and the single steps.

The first step in the approach is to raise research questions and provide appropriate contributions to each research question. Based on the motivation and problem statements we identified the following research Questions:

- RQ1 How can we identify compromised user accounts, based on a user's clickstreams and taking into account a temporal aspect?
- RQ2 How many clicks and background knowledge about the user do we need in order to be able to give a sufficient confidence about the compromising.
- RQ3 Does our approach work for every clickstream data or is it limited to a certain domain?

The first research questions tackles the aspect of identifying fraudulences in a user's account. The methods we can use for these and the following research questions will be explained in more detail in the next section. The second research questions is about providing a confidence for the predictions made by RQ1. More clicks and background knowledge about the user probably leads to an improved confidence of the prediction. The question is, however, how many clicks and background knowledge are needed to get a certain confidence. In addition, the question of which methods are suitable when this critical amount of information is not available to make a statement with a certain confidence is also of interest. The last research questions tackles the applicability of our approach regarding its suitability for certain domains. Here we will examine whether the clickstreams of the different scenarios differ and whether our approach is limited to a specific use case. For each research question will corresponding contributions be provided. After we have sharpened the research questions, we have to identify the relevant literature in the next step. We have already taken a first step here (see section 2 - State of the art). However, other research ideas and applications of methods can lead to further search for relevant literature. This step will therefore be carried out in parallel with the method selection.

In order to evaluate the methods and the approach, data sets are required. Preferably, the selected methods are not tested on a single data set, but on several. Existing datasets such as RecSys Challenge 2015[4] and Yandex[5] all have in common that although they contain clickstreams in different forms, it is unknown whether they contain fraud clicks or not. This means that it is not possible to quantify how good our methods actually are. However, as mentioned in section 2 - State of the art, the topic is related to fraud detection in the financial sector. Therefore, we could consider using financial data sets. There are synthetic data like banksim[6] and paysim[7] available. Therefore, we could take into account to adapt the synthetic financial data sets and considering each financial transaction as a click and subsequent transactions as a clickstream. The transferred amount of money could be considered as the duration on a webpage. The advantage here would be that labeled data would be available. This would facilitate the evaluation, because by comparing with the correct classification it can be determined exactly whether the chosen methods and approach are advantageous to recognize fraudulences in a data set. As a last option, we could create our own data set. Test persons could traverse on server-own systems and thus generate data sets including fraudulences.

Once we have selected one or more data sets, we can select appropriate methods and use them to identify compromised or shared accounts. We will distinguish ourselves from existing methods by using a semantic and structure-based analysis instead of a graph-based analysis. We will go into more detail on the selection of methods and evaluation criteria in section 4 - Methodology.

We will continue to implement the agile approach until our chosen methods outperform existing approaches. In the end of the work we will draw conclusions based on the evaluations and experiences, gained during the research. The insights and selected, and possibly extended, methods are described and summarized. Research questions that have been asked at the beginning of the work are answered with the help of the insights. In addition, our work is embedded in the existing research environment to bring the work into a context. This ensures, on the one hand, that the work is

[4] http://recsys.yoochoose.net/challenge.html, last access: 23rd February 2018
[5] https://www.kaggle.com/c/yandex-personalized-web-search-challenge/data, last access: 23rd February 2018
[6] https://www.kaggle.com/ntnu-testimon/banksim1/data, last access: 23rd February 2018
[7] https://www.kaggle.com/ntnu-testimon/paysim1, last access: 23rd February 2018

again delimited from existing ones, synergies with other areas are identified and possible applications in other areas are presented.

4 METHODOLOGY

For the present work we use methods from quantitative research. We use a hybrid approach to identify fraudulences. With the help of machine learning methods, we can quickly identify abnormalities and interrelationships. With this information we can build up a knowledge base to use this knowledge to identify compromised and shared accounts. In addition, we will enrich the knowledge base with practical knowledge. These two methods, build the base of our hybrid approach. We will explain in the following available methods of subsymbolic AI in more detail and then discuss the possible methods of symbolic AI.

As already mentioned, the identification of fraudulences is predominantly a binary classification problem. A variety of methods of subsymbolic AI are available to us for this purpose. First of all, we have to transform the clickstreams into a corresponding model. For this we can use methods from the field of Text Mining like e. g. a co-occurrence matrix, term-frequency matrix, calculate tfidf to identify the relevance of a click or use Doc2Vec. All of these methods have in common that they produce a numerical vector representation of the sessions or clicks. A numerical representation is recommended for further processing, as it allows us to use elementary algebra and geometry instruments for further processing. A further approach to determine a numerical representation is the use of Latent Semantic Analysis (LSA). We will consider including and identifying semantic information in clickstreams. LSA is used very successfully in the area of Natural Language Processing (NLP). We have already made first attempts to apply LSA to clickstreams.

Besides the appropriate representation of clickstreams, the selection of suitable methods is also crucial. Support Vector machines, Bayesian networks or artificial neural networks could be used. The methods have in common, however, that they consider the data statically and do not take any temporal aspect into account[8]. This has to be taken into account, when representing clickstreams as well as when selecting models, since the preferences of a user can change over time and these clicks should not be considered as fraudulences. Therefore, both the model has to be adapted over time and the temporal component has to be taken into account. Methods may need to be adapted to take this into account. For RQ2, statistical methods are used to validate the prediction, on the one hand, to put the prediction in context with probabilities and thus to determine the quality of the prediction, and on the other hand to justify its applicability.

So far, we have only listed methods of supervised learning. However, unsupervised learning methods can also be used. This could be particularly advantageous if little is known about an user. By means of Nearest Neighbor or k-means clustering the user can be assigned to the most similar set. However, if the density of the cluster changes very much as a result of the assignment, it can be assumed that this user contains fraudulences, since the assignment is unfavorable.

The above methods highlight the application of subsymbolic AI. However, as already mentioned above, we choose a hybrid use of subsymbolic and symbolic AI. We would like to make the findings from the above methods available in a knowledge base. This knowledge base is enriched with methods from the field of symbolic AI. Here we want to make it possible to enrich the knowledge of machine learning with practical knowledge by domain experts. We hope that the hybrid approach will enable us to achieve a higher level of meaningfulness. One difficulty here is to find a suitable representation of both, the insights of machine learning and of practical knowledge and to model them accordingly. We also need to consider a temporal aspect when modeling knowledge. Similarly, the opinions and experience of different domain experts must also be taken into account in the modeling process. Different experts may have different experiences with compromised accounts and provide different experiences.

When selecting the evaluation criteria, it should be noted that in our case it is predominantly a binary classification problem. When identifying the person of a shared account, there could be several classes. In addition, it can be assumed that the considered data sets are very unbalanced. This means that there are a lot of sessions in the clickstream data sets in a certain class. In our case, the extent of sessions that are not compromised or shared will be greatly increased. For unbalanced data, *Area under the Curve (ROC AUC)* is recommended as evaluation criterion. ROC AUC is not reflected by data imbalance. ROC AUC computes the area under the receiver operating characteristic curve, which illustrates the true positive rate (TPR) against the false positive rate (FPR). The fact of ROC AUC being insensitive to class balance makes this evaluation criterion very suitable for our case. In order to compare our methodology and approach with existing methods, we will apply existing methods to the data sets in order to compare them with our methods. In the implementation of existing methods and the subsequent comparisons, the data set and the different methods must always be taken into account. Existing methods often use graph-based methods to identify compromised accounts. These graphs may not be available to us.

5 RESULTS

Currently, the work is in an early stage. The approach of the work was clearly written down, the first related works were identified and possible methods of application were presented. However, during the work the overview of the related work will be enlarged.

Initial work was done to enable a semantic representation of clickstreams. The RecSys Challenge 2015[9] dataset was used for this. The data was represented as a session item matrix. Thus we have created an embedding for each item and session respectively. LSA was then applied to it. LSA allows for representing a session-item matrix A as the product of three matrices: $A = U \cdot \Sigma \cdot V^T$. Here U represents the items and V^T the sessions. With this representation, the items could be assigned to the correct category with a high degree of accuracy. We used Support Vector Machines (SVM) to classify the items in their categories. The assignment to the category was used because this information was available. It was not possible to assign the sessions to users, since no information was known about them. A similarity analysis would have been possible, but could not have been compared with a gold standard. We showed

[8] Except for some NN like recurrent neural networks

[9] http://recsys.yoochoose.net/challenge.html, last access: 23rd February 2018

in this work that LSA is applicable to assign items to their correct category, based on the clickstreams of users. With the help of LSA we exploited the semantics of the clickstreams. In addition we tried to rebuild the taxonomy of the product category by exploiting the semantics in the clickstreams. However, we did not have the taxonomy available. We build it based on the information of shared items for each category. This work helped as a first step for session modeling in clickstreams.

6 CONCLUSIONS AND FUTURE WORK

The main goal of this work is to identify fraudulences and shared user accounts, based on clickstream data. One aspect, besides the identification of fraudulences, is the amount of information, needed to make meaningful predictions. Another aspect that is tackled during this work is the applicability of our approach to different domains. The approach should be abstract enough to be applicable to different clickstream data like e.g. social networks, web shops and web solutions. Most of the related work focus on graph-based solutions to identify hijacked accounts. We will focus on compromised accounts and use clickstream data with respect to a temporal aspect.

We use a hybrid approach, consisting of methods from subsymbolic and symbolic AI. Thus methods from data representation and machine learning are used, as well as logic-based methods. Insights gained from machine learning algorithms and practical knowledge from domain experts are considered in our approach. We hope that this will lead to improved results. Future work includes to identify one or more suitable data sets. The currently available data sets do not fulfill all specified requirements. We are currently considering using one of the synthetic financial data sets as a clickstream dataset, since it contains labeled data. This makes the evaluation credible, rather than data sets that are not known to contain fraudulences. It is important to note that the data set should have different sessions for one user in order to consider the change of preference for one user over time. Besides choosing appropriate data sets for the evaluation, a suitable model for the knowledge base has to be considered. Since we want to store the information in a knowledge base and enrich it with practical knowledge, we need to provide a data model that makes both possible. In addition, this data model must be able to deal with vague and contradictory practical knowledge and allow for a change over time. We can then convert the data into the appropriate data model and apply the selected methods to it. Due to the agile approach we will quickly adapt the results of the first experiments to reflect back the data model as well as the methods used. At the end of the work, we will collect the gained knowledge and thus answer the research questions.

REFERENCES

[1] Palakorn Achananuparp, Hyoil Han, Olfa Nasraoui, and Roberta Johnson. 2007. Semantically Enhanced User Modeling. In *Proceedings of the 2007 ACM Symposium on Applied Computing (SAC '07)*. ACM, New York, NY, USA, 1335–1339.

[2] Juan JosÃĺ GarcÃŋa Adeva and Juan Manuel Pikatza Atxa. 2007. Intrusion detection in web applications using text mining. *Engineering Applications of Artificial Intelligence* 20, 4 (2007), 555 – 566.

[3] Maristella Agosti, Franco Crivellari, and Giorgio Maria Di Nunzio. 2012. Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. *Data Mining and Knowledge Discovery* 24, 3 (01 May 2012), 663–696.

[4] R. Brause, T. Langsdorf, and M. Hepp. 1999. Neural data mining for credit card fraud detection. In *Proceedings 11th International Conference on Tools with Artificial Intelligence*. 103–106.

[5] Philip K. Chan and Salvatore J. Stolfo. 1998. Toward Scalable Learning with Non-uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*. AAAI Press, 164–168.

[6] J. R. Dorronsoro, F. Ginel, C. Sgnchez, and C. S. Cruz. 1997. Neural fraud detection in credit card operations. *IEEE Transactions on Neural Networks* 8, 4 (Jul 1997), 827–834.

[7] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna. 2017. Towards Detecting Compromised Accounts on Social Networks. *IEEE Transactions on Dependable and Secure Computing* 14, 4 (July 2017), 447–460.

[8] Manuel Egele, Gianluca Stringhini, Christopher Krügel, and Giovanni Vigna. 2013. COMPA: Detecting Compromised Accounts on Social Networks. In *NDSS*.

[9] Andrew Foss, Weinan Wang, and Osmar R. ZaÃŕane. 2001. A Non-Parametric Approach to Web Log Analysis. (2001).

[10] S. Ghosh and D. L. Reilly. 1994. Credit card fraud detection with a neural-network. In *1994 Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences*, Vol. 3. 621–630.

[11] Nicholas A. Heard, David J. Weston, Kiriaki Platanioti, and David J. Hand. 2010. Bayesian anomaly detection methods for social networks. *Ann. Appl. Stat.* 4, 2 (06 2010), 645–662.

[12] Terran Lane and Carla E. Brodley. 1997. An Application of Machine Learning to Anomaly Detection. In *In Proceedings of the 20th National Information Systems Security Conference*. 366–380.

[13] Ahmed Metwally, Divyakant Agrawal, and Amr El Abbadi. 2005. Using Association Rules for Fraud Detection in Web Advertising Networks. In *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB '05)*. VLDB Endowment, 169–180.

[14] Bamshad Mobasher. 2005. Web usage mining. In *Encyclopedia of data warehousing and mining*. IGI Global, 1216–1220.

[15] Alan L. Montgomery, Shibo Li, Kannan Srinivasan, and John C. Liechty. 2004. Modeling Online Browsing and Path Analysis Using Clickstream Data. *Marketing Science* 23, 4 (2004), 579–595.

[16] O. Nasraoui and C. Petenes. 2003. An intelligent Web recommendation engine based on fuzzy approximate reasoning. In *Fuzzy Systems, 2003. FUZZ '03. The 12th IEEE International Conference on*, Vol. 2. 1116–1121 vol.2.

[17] Caleb C. Noble and Diane J. Cook. 2003. Graph-based Anomaly Detection. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03)*. ACM, New York, NY, USA, 631–636.

[18] Rainer Olbrich and Christian Holsing. 2011. Modeling Consumer Purchasing Behavior in Social Shopping Communities with Clickstream Data. *International Journal of Electronic Commerce* 16, 2 (2011), 15–40.

[19] Shigeru Oyanagi, Kazuto Kubota, and Akihiko Nakase. 2001. Application of matrix clustering to web log analysis and access prediction. In *in: WEBKDD 2001âĂŤMining Web Log Data Across All Customers Touch Points, Third International Workshop*. 13–21.

[20] Soyeon Park, Joon Ho Lee, and Hee Jin Bae. 2005. End user searching: A Web log analysis of NAVER, a Korean Web search engine. *Library & Information Science Research* 27, 2 (2005), 203 – 221.

[21] Animesh Patcha and Jung-Min Park. 2007. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks* 51, 12 (2007), 3448 – 3470.

[22] Jake Ryan, Meng-Jang Lin, and Risto Miikkulainen. 1998. Intrusion Detection With Neural Networks. In *Advances in Neural Information Processing Systems 10*, Michael I. Jordan, Michael J. Kearns, and Sara A. Solla (Eds.). Cambridge, MA: MIT Press, 943–949.

[23] Narayanan Sadagopan and Jie Li. 2008. Characterizing Typical and Atypical User Sessions in Clickstreams. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*. ACM, New York, NY, USA, 885–894.

[24] David Savage, Xiuzhen Zhang, Xinghuo Yu, Pauline Chou, and Qingmai Wang. 2014. Anomaly detection in online social networks. *Social Networks* 39, Supplement C (2014), 62 – 70.

[25] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. 1999. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum* 33, 1 (Sept. 1999), 6–12.

[26] S. Wang. 2010. A Comprehensive Survey of Data Mining-Based Accounting-Fraud Detection Research. In *2010 International Conference on Intelligent Computation Technology and Automation*, Vol. 1. 50–53.

[27] L. Zhang and Y. Guan. 2008. Detecting Click Fraud in Pay-Per-Click Streams of Online Advertising Networks. In *2008 The 28th International Conference on Distributed Computing Systems*. 77–84.