Department of Computer Science

IT-Security Seminar (SS2019)

Security and Privacy in Social Networks

Philipp Normann — **its103541@fh-wedel.de**

supervised by Gerd Beuster — **gb@fh-wedel.de**

June 12, 2019

Contents

List of Figures

1 Introduction

Online Social Networks (OSNs), such as Twitter, Facebook or Instagram have become increasingly popular over the past decades. As these platforms rose in popularity and their influence on society increased, attackers have also started considering how to use them for malicious activities. The performed attacks range from bots influencing the public discourse to targeted spam and private information being disclosed to unauthorized third parties [Sat+14]. These new risks have resulted in a new research field, studying the nature of these social networks and methods how security and privacy can be ensured in them. In the following sections expected problems, research gaps and possible example problems will be explored. This seminar is based on the chapter *Social Networks* from the Red Book — A Roadmap for Systems Security Research [MB13], but also supplements the topics with recent findings and developments.

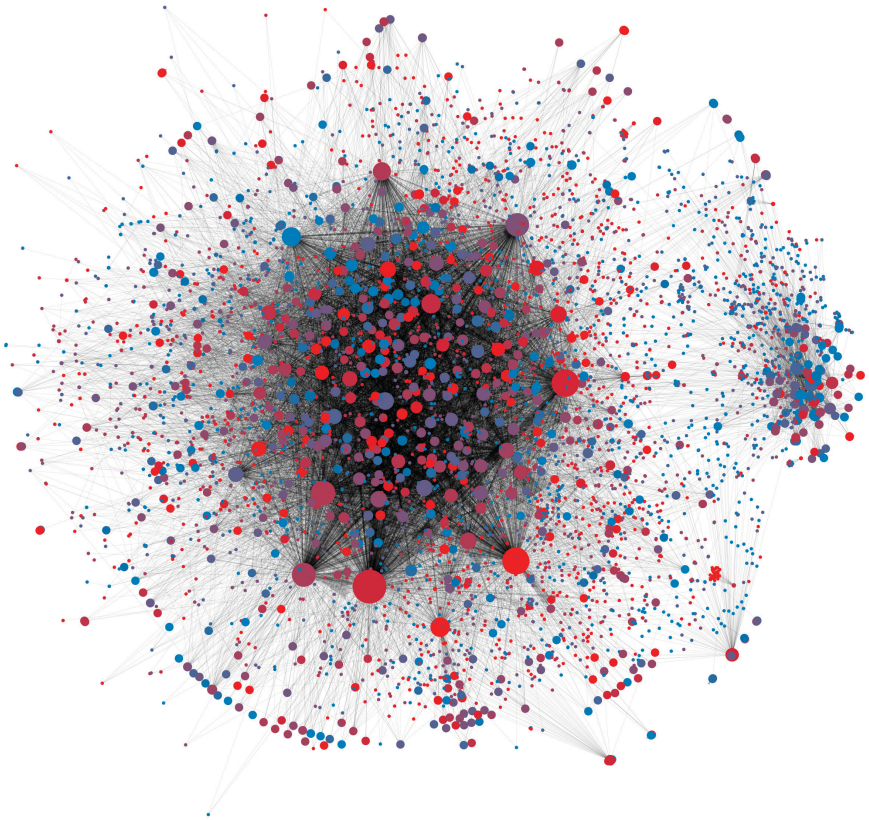

Figure 1: Visualization of a retweet graph, showing how bots (red) are used to influence a political debate of legitimate users (blue) on Twitter [Fer+16]

2 Expected Problems

We are already seeing various attacks being perpetrated on social media platforms such as Facebook, Twitter or Instagram. By befriending strangers, cyber criminals gain access to personal data to perform identity theft or send victims malicious messages including spam, phishing or malware. This section will explore the various security and privacy problems, inherent to the current generation of centralized OSNs, which were presented in the Red book [MB13].

2.1 Privacy

The rising popularity of OSNs has accelerated the appearance of vast amounts of personal information on the internet. These include education, occupation, relationship status, current location, and personal habits. All of this information can be used to launch advanced targeted attacks against people. The online identities can also be linked to offsite behavior due to information which is leaked via OSN integration or third party apps [KW09]. It has also been shown that social bots can successfully be used to infiltrate a social network undetected in order to obtain access to profile information which is not publicly available [Bos+11]. Not only is the consciously shared information exposed, but also other traits of a person, such as sexual orientation, ethnicity, religious and political views, personality traits, intelligence or happiness can be inferred from easily accessible information such as Facebook Likes [KSG13]. The accuracy of such computer-based personality judgment is surprisingly accurate and even surpasses the accuracy of human judgments made by Facebook friends [YKS15], as seen in Figure 2. This vast leakage of private information without the deliberate consent of the users, poses a threat to the trustworthiness of these online platforms and needs to be tackled. Privacy protecting laws and technologies combined with a more conscious user behavior, when it comes to sharing information online, could help to alleviate this problem.

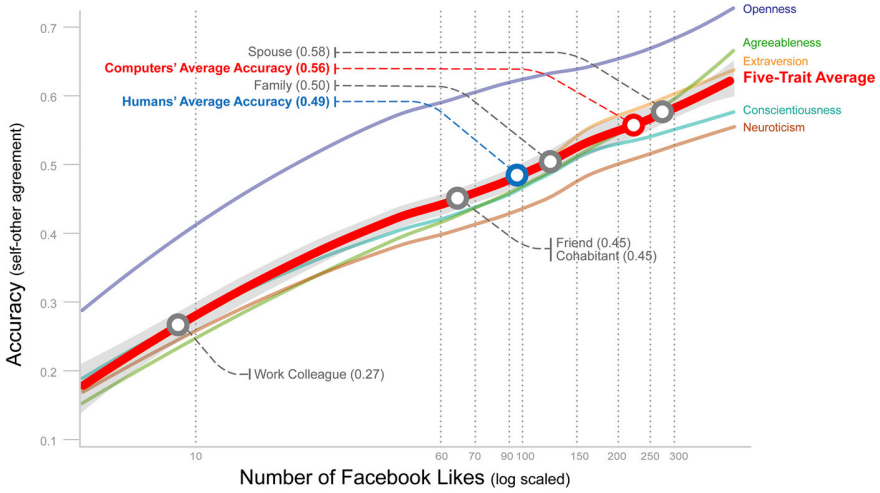

Figure 2: Computer-based personality judgment accuracy of personality traits compared to human performance of different types of relationships. Note that the average computer accuracy is significantly better than that of an average human judge and comparable with an average spouse [YKS15]

2.2 Spam

Online content-sharing platforms, such as OSNs, have become one of the main distribution channels for spammers to spread their malicious messages. Using the personal information, that users provide in their profiles, these spammers can easily personalize their messages in order to increase their success rate. A prominent example for such a campaign resulted in the propagation of a malware called Koobface which attempted to steal sensitive information and form a peer-to-peer botnet [BCF09].

2.3 Sybil Attacks

Multiple generated (Sybil) identities in social networks can be used to outvote honest users, influence online ratings, and manipulate search results. These fake accounts can also be used to manipulate a public debate or censor specific topics by means of hashtag flooding [TGP12]. The sale of such fake accounts is starting develop its own economy in which fake or hijacked accounts are sold in bulk of thousands [Tho+13]. The effects of such Sybil attacks have also been observed on Twitter during the 2019 EU election, where 12 per cent of tweets using hashtags promoted by far-right parties showed clear signs of full automation [Bev19]. In order to mitigate these types of attacks various methods have been proposed, for example clustering user behavior based on click stream data [Wan+13] or applying graph mining techniques to social graphs [Vis+11].

2.4 Authentication

In the age of an ever increasing amount of leaked credentials and a continuous increase in computational power, that can be used to brute force accounts, using only passwords for authentication is insufficient and should be complemented by a second factor. In 2011, in an effort to combat this problem, Facebook introduced Social authentication (SA). Their implementation is based on recognizing friends in pictures. A group of researchers showed that by infiltrating a users social circle and utilizing face recognition software this method of authentication can reliably be bypassed and is therefore also considered insecure [Pol+12]. Second factors such as OTPs, e.g. using hardware tokens or via SMS, are better suited for securing online accounts. It should be noted that, although requiring more preparations and technical expertise, compared to a regular phishing campaign, a real-time MITM phishing attack is still possible, even when OTPs are enabled as a second factor.

2.5 Third Parties

The integration of third party apps on social media sites and vise versa can also pose a security and privacy risk to users of OSNs. Not only can the browsing behavior of the users be tracked in unwanted ways, but giving permissions to a third party app can also result in unwanted disclosure of private information. In 2018 a whistle-blower and former employee of a company called *Cambridge Analytica* exposed a vast leakage of personal data, including private messages and profile information of friends, through a 3rd party app on Facebook. The leaked dataset contained psychological profiles of more than 230 million Americans and was used for targeted advertisement and support of political campaigns, such as the 2016 presidential campaign of Donald Trump [CG18].

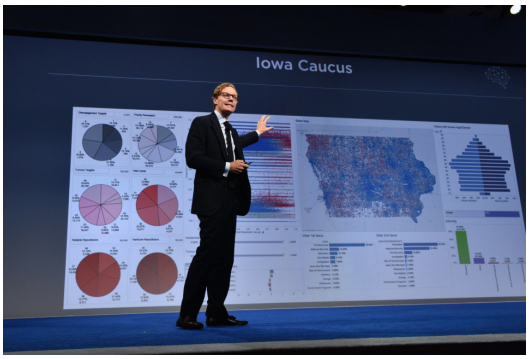

Figure 3: Alexander Nix, former CEO of Cambridge Analytica, presenting a psychological profiling dashboard at the Concordia Summit in New York [MS18]

3 Worst-case Scenario

As already indicated in the previous section, social networks can be used to spread false information [KS18], perform political censorship [TGP12], bias public opinion [Fer+16] and also attack single users. These possible ways of exploiting OSNs can have drastic impacts on society and their prevention should therefore be highly prioritized topics for social network platforms such as Facebook, Twitter or Instagram. An even worse scenario can occur if the platform providers themselves start to intentionally manipulate the networks in malicious ways we haven't seen before. A glimpse of what might be possible in this direction can be seen in China, where a social credit system (SCS) is being developed and tested, that rates all its citizens according to their online and offline behavior. This can possibly have drastic implications on the freedom of speech, e.g. if posts criticizing the political party of power get scored negatively [CTS18].

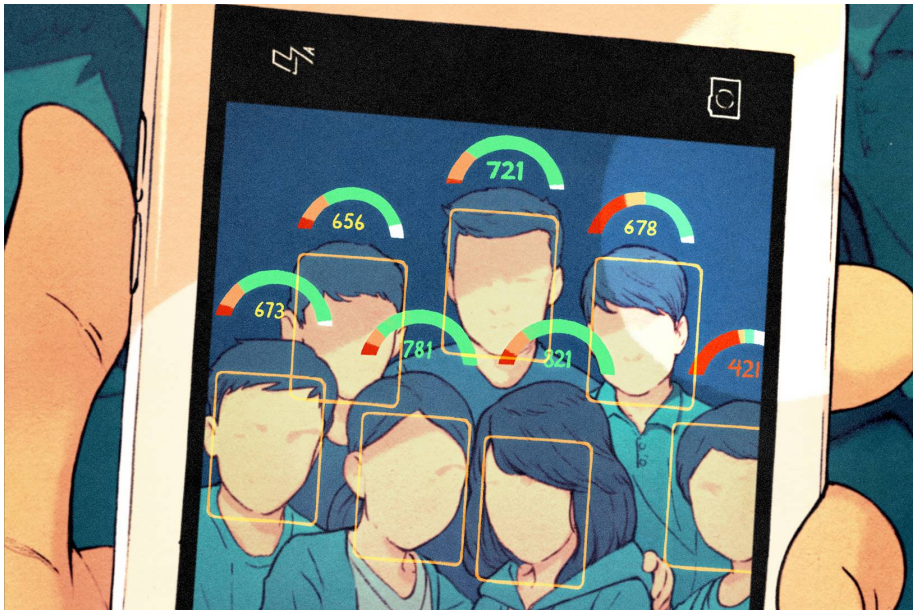

Figure 4: An artistic visualization of the Chinese SCS by Kevin Hong [Kob19]

4 Research Gaps

Following the previously explained, current and expected security and privacy problems that arise in the context of OSNs, the following section will explore the inherent research gaps presented in the Red Book [MB13] and complement them with recent findings and possible solutions.

4.1 Trustworthiness of Information

As mentioned in the previous sections, false information also known as fake news is often spread through the web and especially social media platforms. Several research studies have determined the impact of false information in social networks in terms of user engagement metrics, such as the number of likes, reshares, and pre-removal lifetime. They discovered that some pieces of false information are highly impactful, they are liked, shared, and commented on more, generate deeper cascades of reshares than true information pieces, survive for a long time, and spread across the web effectively [KS18; Zub+16]. This can result in widespread real-world impact on the public opinion. In addition OSNs can produce echo-chambers, which lead to polarization and can further encourage the spread of false information, as seen in Figure 5.

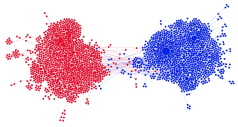

Figure 5: Echo-chambers in a retweet graph of a controversial topic [Gar+17]

Accessing the trustworthiness of information can be very challenging especially in the setting of anonymity or in cases where there is no other way to verify a piece of content. Several methods of automatically identifying false information have been proposed, they can be categorized into two major categories: feature engineering based and propagation based. Whereas the feature engineering based models create features from textual properties [Pér+17] and their relation to other existing information, the propagation based methods try to model how true information propagates in these networks in order to detect anomalies of these models as false information [KS18]. Some research has also been conducted on leveraging the wisdom of the crowd to detect and reduce the spread of misinformation [Kim+18].

Various algorithms have already been developed for detection of false information in different domains. However, they are not directly comparable to each other due to the lack of large-scale publicly available datasets. This prevents a benchmark comparison between different categories of algorithms. Some recent datasets, such as LIAR [Wan17] have been created but standardized comparison of existing algorithms on these datasets has yet to be conducted [KS18]. Thus, the estimation of the trustworthiness for different information sources still remains an interesting research gap, including many subtopics such as the problem of bridging echo chamber, detecting false multimedia content, produced using deep learning methods such as [Zak+19] or automatically fact checking information from a knowledge base.

It should be noted that the distribution of fake news is often heavily supported using social bots or so called sockpuppets. Therefore the solution to this problem is also closely related to the topic of detecting fake identities in social graphs [Fer+16], see subsection 5.3 for more details.

4.2 Real-time Data Processing

Another challenge in regards to protecting OSNs from malicious activities, is the sheer amount of ongoing events that need to be processed by such a real-time system. This can not be solved using traditional data mining techniques, but is rather a stream processing problem. The main challenge is the identification of malicious or fraudulent sources in real-time. A published approach towards this goal is the so called *Facebook Immune System*, which performs real-time checks and classifications on every read and write action [SCM11]. Although a step in the right direction, other researchers have shown that the security defenses, at least the deployed solution at Facebook in 2011, was not effective enough in detecting or stopping large-scale infiltration, as it occurred in practice [Bos+11].

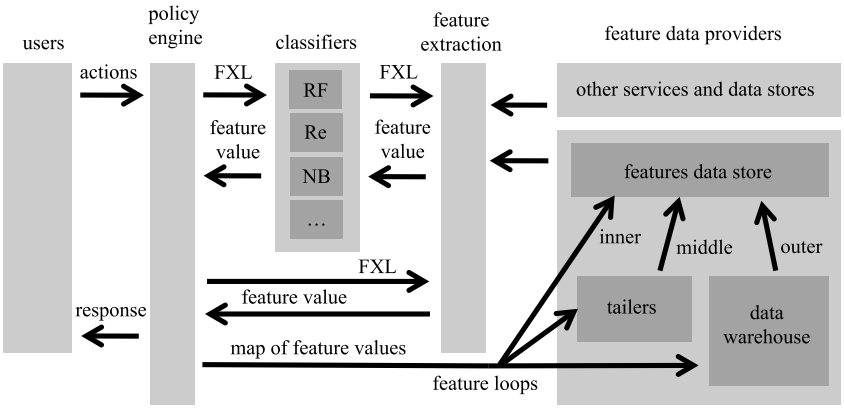

Figure 6: High-level overview of the Facebook Immune System [SCM11]

4.3 Anomaly Detection in Social Graphs

There have already been numerous studies, such as [Sha17] and [Vis+14], covering the application of graph mining algorithms to the problem of identifying anomalies in the structures of social graphs. These methods can be successfully used to detect possible attacks, in which attackers do not conform to the expected social behavior. However, since defenders and attackers are in a cat-and-mouse game in which both continuously improve their methods, this research area still remains relevant and offers plentiful directions for future work [Sha17].

4.4 Coping with the Dynamicity of Social Graph Data

In addition to the problem of detecting anomalies in snapshots of social graph data, tracking changes over time can give an even better insight into the dynamic nature of the network and attackers behavior. One method for pattern learning and anomaly detection in streams of graph data, promising to fulfill this need, is called PLADS and was published in 2015 by Eberle and Holder [EH15]. Another timing based approach for detecting accounts that act in loose synchrocy is called SyncroTrap and is actively used by Facebook and Instagram [Cao+14].

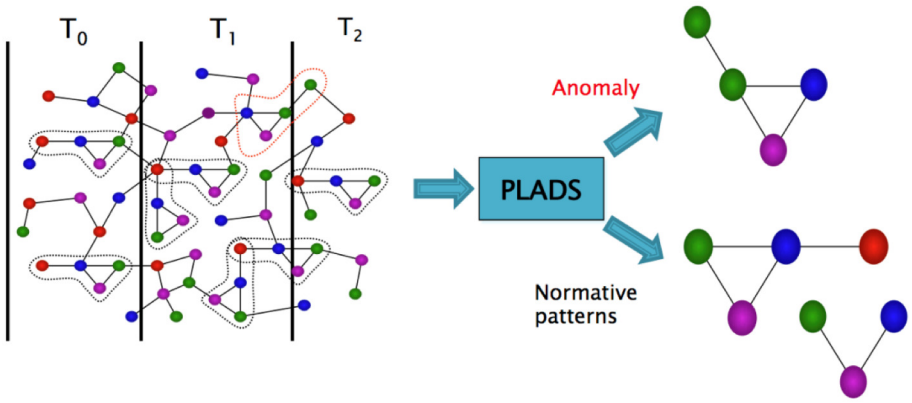

Figure 7: Information about entities and relationships stream in over time, and PLADS detects anomalies in the graph [EH15]

4.5 Security and Privacy Trade-Off

In the analysis of social interaction data for the purpose of detecting cyber criminals, there is always a trade-off between privacy and security. Data collection and processing can result in the detection of malicious accounts, but in order to be practical this also has to be done in a privacy preserving manner. This is especially difficult for graph data, as only removing personal identifiable information from the nodes is not enough, as the network and other meta data can be used to deanonymize nodes. Although multiple studies, covering the topic of effectively anonymizing graph datasets have been conducted [LT08; FNT08; ZCÖ09], the topic is still quite new compared to the topic of ananomizing tabular datasets. For example, it is still not well defined, what kind of attacks on anonymity these procedures protect against.

5 Example Problems

The following section will explore a couple of research problems in the field of securing social networks, that are mentioned in the Red Book [MB13]. The problems will be explained and recent research that has been conducted in the respective directions will be mentioned.

5.1 Measure of Truthfulness

How can we build a system that is able to reliably measure the truthfulness of information that is consumed on social media? This question has been the topic of multiple studies, as discussed in subsection 4.1. A recent paper compared seven state-of-the-art hate speech detection models from prior work, and showed that they perform well only when tested on the same type of data they were trained on and showed that all proposed detection techniques are brittle against adversaries who can (automatically) insert typos, change word boundaries or add innocuous words to the original hate speech. They suggest that using character-level features instead of word-level features would make the textual models systematically more resistant to adversarial attacks [Grö+18].

5.2 Real-time Detection of Cyber Criminals

How can we build a system that is able to detect cyber criminals in real time? Such a system requires massive parallelization and distribution in order to be able to process the large amounts of data that are created in OSNs. This question has been the topic of multiple studies, as discussed in subsection 4.2. Another recent paper proposed a timing based detection mechanism, called SynchroTrap, that can uncover large groups of malicious accounts that act in loose synchronicity. This system was also successfully deployed at Facebook and Instagram and uncovered more than two million malicious accounts involved in large attack campaigns within one month [Cao+14].

5.3 Identification of Fake Identities

How could we build a system that could identify fake profiles? In addition to the techniques of graph mining discussed in subsection 4.3, other means of Sybil account detection can be used. A study, analyzing Sybil accounts on an OSN called Renren, shows that a threshold-based classifier is sufficient to catch 99% of Sybils, with low false positive and negative rates [Yan+14].

6 Closing Remarks

In this paper, I gave a few examples for how the security and privacy in online social networks is currently endangered, presented an overview of the types of problems that arise and also highlighted possible solutions and related research areas. The presented challenges are real and require social network platforms, such as Facebook, Twitter or Instagram to provide solutions to these problems in a timely manner, if they do not want to risk a loss of trustworthiness and users.

However in my opinion the currently prevalent social networks, due to their centralized nature, pose a privacy risk in themselves. So perhaps a more decentralized federated network, similar to Mastodon [Mas16], that utilizes privacy preserving machine learning methods to protect the network from spam and false information and gives the users full control and transparency over the use of their personal data, could be developed as a viable alternative for the future.

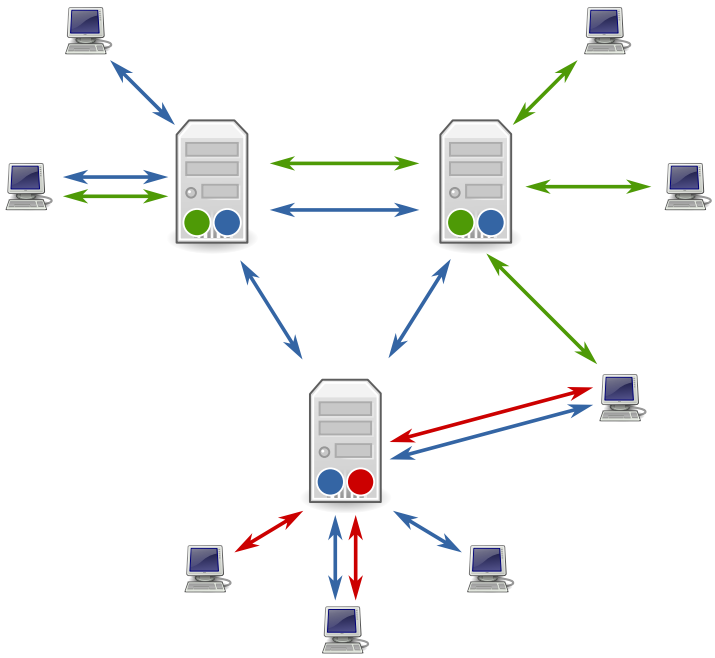

Figure 8: A diagram of servers and clients in a federated network, an alternative to centralized architectures, prevalent in current social networks [Esh07]

References

[BCF09] Jonell Baltazar, Joey Costoya, and Ryan Flores. "The real face of koobface: The largest web 2.0 botnet explained". In: *Trend Micro Research* 5.9 (2009), p. 10. URL: http://www.trendmicro.com.br/cloud-content/us/pdfs/security-intelligence/white-papers/wp_the-real-face-of-koobface.pdf.

[Bev19] Emmi Bevensee. *Opinion: We tracked European election bots – what we found should scare you*. May 2019. URL: https://www.independent.co.uk/voices/european-elections-parliament-bots-social-media-matteo-salvini-far-right-a8924831.html (visited on 06/12/2019).

[Bos+11] Yazan Boshmaf et al. "The socialbot network: when bots socialize for fame and money". In: *Proceedings of the 27th annual computer security applications conference*. ACM. 2011, pp. 93–102. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.661.5915&rep=rep1&type=pdf.

[CG18] Carole Cadwalladr and E Graham-Harrison. "The Cambridge analytica files". In: *The Guardian* 21 (2018), pp. 6–7. URL: https://davelevy.info/Downloads/cabridgeananalyticafiles%20-theguardian_20180318.pdf.

[Cao+14] Qiang Cao et al. "Uncovering large groups of active malicious accounts in online social networks". In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM. 2014, pp. 477–488. URL: https://www.eecis.udel.edu/~ruizhang/CISC859/S17/Paper/p19.pdf.

[CTS18] Martin Chorzempa, Paul Triolo, and Samm Sacks. "18-14 China's Social Credit System: A Mark of Progress or a Threat to Privacy?" In: (2018). URL: https://piie.com/system/files/documents/pb18-14.pdf.

[EH15] William Eberle and Lawrence Holder. "Scalable anomaly detection in graphs". In: *Intelligent Data Analysis* 19.1 (2015), pp. 57–74. URL: http://ailab.wsu.edu/adgs/pdfs/EberleIDA2015wm.pdf.

[Esh07] Benjamin D. Esham. *A diagram of some Usenet servers and clients*. Aug. 2007. URL: https://commons.wikimedia.org/wiki/File:Usenet_servers_and_clients.svg (visited on 06/12/2019).

[FNT08] Tomás Feder, Shubha U Nabar, and Evimaria Terzi. "Anonymizing graphs". In: *arXiv preprint arXiv:0810.5578* (2008). URL: https://arxiv.org/pdf/0810.5578.

[Fer+16] Emilio Ferrara et al. "The rise of social bots". In: *Communications of the ACM* 59.7 (2016), pp. 96–104. URL: https://dl.acm.org/ft_gateway.cfm?id=2818717&type=pdf.

[Gar+17] Kiran Garimella et al. “Reducing controversy by connecting opposing views”. In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM. 2017, pp. 81–90. URL: https://arxiv.org/pdf/1611.00172.

[Grö+18] Tommi Gröndahl et al. “All You Need Is ” Love ” : Evading Hate Speech Detection”. In: 2018. URL: https://pdfs.semanticscholar.org/2e4e/cef8d9295a6913f458a9aef290b3e67de060.pdf.

[Kim+18] Jooyeon Kim et al. “Leveraging the crowd to detect and reduce the spread of fake news and misinformation”. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM. 2018, pp. 324–332. URL: https://people.mpi-sws.org/~manuelgr/pubs/reviewers-misinformation.pdf.

[Kob19] Nicole Kobie. *The complicated truth about China's social credit system*. June 2019. URL: https://www.wired.co.uk/article/china-social-credit-system-explained (visited on 06/12/2019).

[KSG13] Michal Kosinski, David Stillwell, and Thore Graepel. “Private traits and attributes are predictable from digital records of human behavior”. In: *Proceedings of the National Academy of Sciences* 110.15 (2013), pp. 5802–5805. URL: https://www.pnas.org/content/pnas/110/15/5802.full%5C%5C.

[KW09] Balachander Krishnamurthy and Craig E Wills. “On the leakage of personally identifiable information via online social networks”. In: *Proceedings of the 2nd ACM workshop on Online social networks*. ACM. 2009, pp. 7–12. URL: https://www.ftc.gov/sites/default/files/documents/public_comments/privacy-roundtables-comment-project-no.p095416-544506-00010/544506-00010.pdf.

[KS18] Srijan Kumar and Neil Shah. “False information on web and social media: A survey”. In: *arXiv preprint arXiv:1804.08559* (2018). URL: https://arxiv.org/pdf/1804.08559.

[LT08] Kun Liu and Evimaria Terzi. “Towards identity anonymization on graphs”. In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM. 2008, pp. 93–106. URL: http://www.csd.uoc.gr/~hy558/papers/graph_anonymity.pdf.

[MB13] Evangelos Markatos and Davide Balzarotti, eds. *The Red Book: A Roadmap for Systems Security Research*. The SysSec Consortium, Aug. 2013. URL: http://www.red-book.eu/m/documents/syssec_red_book.pdf.

[Mas16] Mastodon. *Your self-hosted, globally interconnected microblogging community*. Feb. 2016. URL: https://github.com/tootsuite/mastodon (visited on 06/12/2019).

[MS18] Phil McCausland and Anna Schecter. *Trump-linked consultants harvested data from millions on Facebook*. Mar. 2018. URL: https://www.nbcnews.com/news/us-news/cambridge-analytica-harvested-data-millions-unsuspecting-facebook-users-n857591.

[Pér+17] Verónica Pérez-Rosas et al. "Automatic detection of fake news". In: *arXiv preprint arXiv:1708.07104* (2017). URL: https://arxiv.org/pdf/1708.07104.

[Pol+12] Iasonas Polakis et al. "All your face are belong to us: breaking Facebook's social authentication". In: *Proceedings of the 28th Annual Computer Security Applications Conference*. ACM. 2012, pp. 399–408. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.926.2019&rep=rep1&type=pdf.

[Sat+14] Sudarshan Kudlur Satyanarayana et al. "Security and Privacy in Online Social Networks: A Survey." In: *EAI Endorsed Trans. Indust. Netw. & Intellig. Syst.* 1.1 (2014), e3. URL: https://pdfs.semanticscholar.org/165b/942b6ff24f35c04b08eaac5904afc2ef6bc7.pdf.

[Sha17] Neil Shah. "Anomaly Detection in Large Social Graphs". Carnegie Mellon University, October 2017. URL: http://reports-archive.adm.cs.cmu.edu/anon/2017/CMU-CS-17-123.pdf.

[SCM11] Tao Stein, Erdong Chen, and Karan Mangla. "Facebook immune system". In: *Proceedings of the 4th workshop on social network systems*. ACM. 2011, p. 8. URL: http://css.csail.mit.edu/6.858/2014/readings/facebook-immune.pdf.

[TGP12] Kurt Thomas, Chris Grier, and Vern Paxson. "Adapting social spam infrastructure for political censorship". In: *Presented as part of the 5th {USENIX} Workshop on Large-Scale Exploits and Emergent Threats*. 2012. URL: https://www.usenix.org/system/files/conference/leet12/leet12-final13_0.pdf.

[Tho+13] Kurt Thomas et al. "Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse". In: *Presented as part of the 22nd {USENIX} Security Symposium ({USENIX} Security 13)*. 2013, pp. 195–210. URL: https://www.usenix.org/system/files/conference/usenixsecurity13/sec13-paper_thomas.pdf.

[Vis+11] Bimal Viswanath et al. "An analysis of social network-based sybil defenses". In: *ACM SIGCOMM Computer Communication Review* 41.4 (2011), pp. 363–374. URL: https://www.researchgate.net/profile/Krishna_P_Gummadi/publication/221164246_An_Analysis_of_Social_Network-Based_Sybil_Defenses/links/55b4c5a908ae092e9655685d.pdf.

[Vis+14] Bimal Viswanath et al. "Towards Detecting Anomalous User Behavior in Online Social Networks". In: *USENIX Security Symposium*. 2014. URL: https://pdfs.semanticscholar.org/f0d8/30b12937755c75b7e07961fd0f509bead7a8.pdf.

[Wan+13] Gang Wang et al. "You Are How You Click: Clickstream Analysis for Sybil Detection". In: *USENIX Security Symposium*. 2013. URL: https://pdfs.semanticscholar.org/22ba/0d428dc3935bb466ef5ae6414473b86327b0.pdf.

[Wan17] William Yang Wang. ""liar, liar pants on fire": A new benchmark dataset for fake news detection". In: *arXiv preprint arXiv:1705.00648* (2017). URL: https://arxiv.org/pdf/1705.00648.

[Yan+14] Zhi Yang et al. "Uncovering social network sybils in the wild". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 8.1 (2014), p. 2. URL: https://arxiv.org/pdf/1106.5321.

[YKS15] Wu Youyou, Michal Kosinski, and David Stillwell. "Computer-based personality judgments are more accurate than those made by humans". In: *Proceedings of the National Academy of Sciences* 112.4 (2015), pp. 1036–1040. URL: https://www.pnas.org/content/pnas/112/4/1036.full.pdf.

[Zak+19] Egor Zakharov et al. "Few-Shot Adversarial Learning of Realistic Neural Talking Head Models". In: *CoRR* abs/1905.08233 (2019). URL: http://arxiv.org/pdf/1905.08233.

[ZCÖ09] Lei Zou, Lei Chen, and M Tamer Özsu. "K-automorphism: A general framework for privacy preserving network publication". In: *Proceedings of the VLDB Endowment* 2.1 (2009), pp. 946–957. URL: https://cs.uwaterloo.ca/~tozsu/ddbms/publications/other/vldb09-privacy.pdf.

[Zub+16] Arkaitz Zubiaga et al. "Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads". In: *PloS one*. 2016. URL: https://arxiv.org/pdf/1511.07487.