

# Cybersafety in Modern Online Social Networks

Edited by

Jeremy Blackburn<sup>1</sup>, Emiliano De Cristofaro<sup>2</sup>, Michael Sirivianos<sup>3</sup>,  
and Thorsten Strufe<sup>4</sup>

1 University of Alabama at Birmingham, US, [blackburn@uab.edu](mailto:blackburn@uab.edu)

2 University College London, GB, [e.decristofaro@ucl.ac.uk](mailto:e.decristofaro@ucl.ac.uk)

3 Cyprus University of Technology – Lemesos, CY, [michael.sirivianos@cut.ac.cy](mailto:michael.sirivianos@cut.ac.cy)

4 TU Dresden, DE, [thorsten.strufe@tu-dresden.de](mailto:thorsten.strufe@tu-dresden.de)

---

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 17372 “Cybersafety in Modern Online Social Networks.” The main motivation behind the seminar stems from the increased relevance of threats and challenges in the context of cybersafety, especially in modern online social networks, where the range of malicious activities perpetrated by malevolent actors is regrettably wide. These include spreading malware and spam, controlling and operating fake/compromised accounts, artificially manipulating the reputation of accounts and pages, and spreading false information as well as terrorist propaganda. The reasons for the success of such attacks are manifold. The users of social networking services tend to extend their trust of the services and profiles of their acquaintances to unknown users and other third parties: despite the service providers’ attempts at keeping their audiences identifiable and accountable, creating a fake profile, also in another person’s name, is very simple. Even partially or fully taking over a profile is comparatively easy, and comes with the benefit of the trust this profile has accrued over time, as many credentials are easy to acquire. Further, even seemingly innocuous issues such as the design and presentation of user interfaces can result in implications for cybersafety. The failure to understand the interfaces and ramifications of certain online actions can lead to extensive over-sharing. Even the limited information of partial profiles may be sufficient for abuse by inference on specific features only. This is especially worrisome for new or younger users of a system that might unknowingly expose information or have unwanted interactions simply due to not fully understanding the platform they are using.

Unfortunately, research in cybersafety has looked at the various sub-problems in isolation, almost exclusively relying on algorithms aimed at detecting malicious accounts that act similarly, or analyzing specific lingual patterns. This ultimately yields a cat-and-mouse game, mostly played on economic grounds, whereby social network operators attempt to make it more and more costly for fraudsters to evade detection, which unfortunately tends to fail to measure and address the impact of safety threats from the point of view of regular individuals. This prompts the need for a multi-faceted, multi-disciplinary, holistic approach to advancing the state of knowledge on cybersafety in online social networks, and the ways in which it can be researched and protected. Ultimately, we want to work towards development of a cutting-edge research agenda and technical roadmap that will allow the community to develop and embed tools to detect malice within the systems themselves, and to design effective ways to enhance their safety online.

This seminar was intended to bring together researchers from synergistic research communities, including experts working on information and system security on one hand, and those with expertise in human/economic/sociological factors of security on the other. More specifically, in the field of cybersafety, there exist a number of interconnected, complex issues that cannot be addressed in isolation, but have to be tackled and countered together. Moreover, it is necessary for these challenges to be studied under a multi-disciplinary light. Consequently, we identified and focused on the most relevant issues in cybersafety, and explored both current and emerging solutions. Specifically, we discussed four problems that are the most pressing both in terms of



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Cybersafety in Modern Online Social Networks, *Dagstuhl Reports*, Vol. 7, Issue 9, pp. 47–61

Editors: Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, and Thorsten Strufe



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

negative impact and potential danger on individuals and society, and challenging open research problems requiring a multi-disciplinary approach: Cyberbullying & Hate Speech, CyberFraud & Scams, Reputation Manipulation & Fake Activities, and Propaganda.

Overall, the seminar was organized to include a number of long talks from senior experts in the field, covering the four main topics above, followed by a series of short talks from the participants about work in progress and recent results, and finally working groups to foster collaborations, brainstorming, and setting of a research agenda forward.

**Seminar** September 10–13, 2017 – <http://www.dagstuhl.de/17372>

**1998 ACM Subject Classification** Human safety, Security

**Keywords and phrases** Cybersafety, Online Social Networks, Security and Privacy, Legal and Ethical Issues on the Web

**Digital Object Identifier** 10.4230/DagRep.7.9.47

**Edited in cooperation with** Savvas Zannettou

## 1 Executive Summary

*Jeremy Blackburn*

*Emiliano De Cristofaro*

*Gianluca Stringhini*

*Michael Sirivianos*

*Thorsten Strufe*

**License** © Creative Commons BY 3.0 Unported license  
© Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Michael Sirivianos, and Thorsten Strufe

The Dagstuhl Seminar 17372 “Cybersafety in Modern Online Social Networks” was a short two and a half day seminar, which took place September 10th–13th, 2017. Its main goal was to bring together researchers from various research areas related to cyberfraud and cybersafety in online social network, and to inspire them to exchange results, practical requirements, and ethical/legal implications related to user-driven research.

**First Day.** The seminar started with a short self-introduction of all the participants, then, we had an initial brainstorming session to identify main topics of interest, various aspects involved in them, and a balance in terms of interdisciplinary representation. Specifically, we focused on scams happening in online social network and hate speech, while paying special attention to the protection of minors. The aspects discussed were related to algorithmic, user, understanding/modeling, ethical, and privacy aspects of working in this line of research.

The brainstorming session concluded with the discussion of the following tangible research directions:

1. We should work on detection, prevention, and mitigation of hate speech.
2. All solutions should be in accordance of regulations.
3. We should pay particular attention to false positives, as users can easily lose their trust in the platform.
4. We should take into consideration the role of proxies, which act as biases on the data.
5. We should focus on counter-terrorism research that aims to distinguish vulnerable population in order to recruit them for propaganda purposes.

We then had four long talks throughout the day. The first speaker, Jeremy Blackburn (University of Alabama at Birmingham, US), described his work on cyberbullying and hate

speech that includes studying behavior on video games as well as fringe Web communities like 4chan. The second speaker, Filippo Menczer (Indiana University – Bloomington, US), presented how misinformation is spread on Twitter. Specifically, he presented how false information as well as the respective fact-checking efforts are diffused on the Twitter network. The third speaker, Gianluca Stringhini (University College London, GB) presented his work on cyberfraud and scams, focusing on deceptive techniques employed by malicious users in order to scam benign users on online dating sites. The last speaker of the first day was Awais Rashid (Lancaster University, GB), who described his work related to child sex offenders and how he coordinated with Police bodies in order to undertake research on this topic. Also, he presented the ethical considerations when doing research with sensitive data, like those used for this study.

**Second Day.** The morning of the second day focused on giving an overview of work done on a variety of topics related to the main topics of the seminar (through short talks from the participants). More specifically, Zinaida Benenson (Universitat Erlangen – Nurnberg, DE), described her work on spear fishing, where malicious users aim to deceive users by sending email that contain malicious URLs. Then, Michael Sirivianos (Cyprus University of Technology- Lemesos, CY) presented his work on how to combat friend spam by analyzing the underlying network of social rejections. The next talk was by Alexandra Olteanu (IBM TJ Watson Research Center – Yorktown Heights, US), who discussed some preliminary results on work done on hate speech. Srijan Kumar (Stanford University, US) then showed how sockpuppet accounts are used in social networks to change and manipulate the opinions of other users of the platform. Savvas Zannettou (Cyprus University of Technology – Lemesos, CY) presented his research on how news propagates across multiple Web communities, and how to measure their influence. Then, Manuel Egele (Boston University, US), presented COMPA, which is a system that captures the behavioral profile of the user in order to identify possible account compromises. Huy Kang Kim (Korea University – Seoul, KR) talked about malicious users exploiting video games to make money. Next, Oana Goga (MPI-SWS – Saarbrücken, DE) described how online identities can be strengthened by combining multiple weak identities. The last talk was by Julien Freudiger (Apple Computer Inc. – Cupertino, US), who covered public privacy and safety guidelines used at Apple.

The afternoon was dedicated to two parallel working groups focused on discussions around a particular topic, specifically, one was about future directions on hate speech research, and another about ethical considerations that researchers should keep in mind when working with users or user data.

**Third Day.** The final day of the seminar had two more parallel working groups, one on research directions related to cyberfraud in online social network, and another on algorithmic biases and possible solutions to avoid it. We then had a discussion summarizing the work and the discussion done in the various working groups and ended up with ideas for future events, collaborations, and follow-ups.

**Acknowledgments.** The organizers of this workshop acknowledge research funding from the European Union’s Horizon 2020 Research and Innovation programme under the Marie Skłodowska-Curie Grant Agreement No 691025. The organizers would like to thank the Schloss Dagstuhl for the professional, productive, and enjoyable atmosphere it provides and for their invaluable support. Finally, we are grateful to and Seda Guerses for taking notes during two of the working groups and to Savvas Zannettou for coordinating the writing of this report and taking notes throughout the seminar.

## 2 Contents

### Executive Summary

<i>Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Michael Sirivianos, and Thorsten Strufe</i> . . . . .	48
---	----

### Overview of Talks

Spear Phishing: Email versus Facebook <i>Zinaida Benenson</i> . . . . .	52
From Pool's Closed to Gas The Kikes <i>Jeremy Blackburn</i> . . . . .	52
Compromised Social Network Accounts – Detection and Incentives <i>Manuel Egele</i> . . . . .	53
Privacy & Safety at Apple <i>Julien Freudiger</i> . . . . .	54
Strengthening Weak Identities Through Inter-Domain Trust Transfer <i>Oana Goga</i> . . . . .	54
Game BOTs (and economic scale of their black money) <i>Huy Kang Kim</i> . . . . .	55
An Army of Me: Sockpuppets in Online Discussion Communities <i>Srijan Kumar</i> . . . . .	55
The spread of misinformation in social media <i>Filippo Menczer</i> . . . . .	56
Hate Speech: Thoughts on the Role of User Aspects & External Events <i>Alexandra Olteanu</i> . . . . .	56
Your words betray you! The role of language in cyber crime investigations <i>Awais Rashid</i> . . . . .	56
Combating Friend Spam Using Social Rejections <i>Michael Sirivianos</i> . . . . .	57
Cyberfraud and Scams <i>Gianluca Stringhini</i> . . . . .	57
The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources <i>Savvas Zannettou</i> . . . . .	58

### Working groups

Algorithmic Biases <i>Seda F. Guerses and Savvas Zannettou</i> . . . . .	58
Ethics <i>Seda F. Guerses and Savvas Zannettou</i> . . . . .	59
Cyberfraud <i>Savvas Zannettou</i> . . . . .	59


Hate Speech	
<i>Savvas Zannettou</i> . . . . .	60
<b>Participants</b> . . . . .	<b>61</b>

### 3 Overview of Talks

In this section, we provide an overview of the talks given at our Dagstuhl seminar, ordered alphabetically, as per the speaker’s last name.

#### 3.1 Spear Phishing: Email versus Facebook

*Zinaida Benenson (Universität Erlangen-Nürnberg, DE)*

License  Creative Commons BY 3.0 Unported license  
© Zinaida Benenson

Joint work of F. Gassmann, A. Girard, N. Hintz, R. Landwirth, A. Luder

Security incidents often start with a click on an infected link or attachment in a spear phishing message. Methods that persuade users to execute the fatal click are getting more and more sophisticated, making defense especially difficult. Drawing from results of an experiment where we sent to over 1600 users an email or a Facebook message with a link to (non-existing) party pictures from a non-existing person, we argue that a carefully targeted and timed message could deceive virtually everyone. Addressing targets by names seems to be especially important via email, resulting in much higher efficiency than messages with non-personalized greetings. On Facebook, however, this targeting technique does not make any difference. The most frequently reported reason for clicking was curiosity (34%), followed by the explanations that the message fit recipient’s expectations (27%). Moreover, 16% thought that they might know the sender. These results show that people’s decisional heuristics are relatively easy to misuse in a targeted attack, making defense especially challenging.

#### 3.2 From Pool’s Closed to Gas The Kikes

*Jeremy Blackburn (University of Alabama at Birmingham, US)*

License  Creative Commons BY 3.0 Unported license  
© Jeremy Blackburn

**Main reference** Gabriel Emile Hine, Jeremiah Onalapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, Jeremy Blackburn: “Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan’s Politically Incorrect Forum and Its Effects on the Web”, in Proc. of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017., pp. 92–101, AAAI Press, 2017.

**URL** <https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15670>

The Internet has allowed for unprecedented flow of information, revolutionized global commerce, and been embraced as a ubiquitous communication channel that enables world wide communities. However, much of what exists now on the Web was not conceived of when it was created, and its pervasive nature has revealed gaps in our understanding, enabling a proliferation of abusive behavior causing real-world harm in ways never really imagined. Cyberbullying, online harassment, and hate speech have quickly reached epidemic proportions. Their effects spill over into the real world, with sometimes violent consequences. While this type of behavior has been studied in offline contexts, its relatively quick growth online has left us with substantial gaps in our knowledge, gaps we must fill if we hope to combat it.

As a first step, we must understand the real-world psychology behind this behavior. For example, substantial literature has shown that the bully label is not one-size-fits-all. Instead, there are multiple types of bullies and harassers, each with unique motivations and views

of their victims. Once we understand the underlying rationale, we can begin exploring the features of the online world that enable the proliferation of this behavior. Here, we must examine how things like relative anonymity and asymmetric communication channels can empower harassers, enable recruitment of organized hate speech campaigns, etc.

While theory and controlled experiments can provide us with insight, they lack empirical evidence. Thus, the next step is to seek out and measure this behavior in the wild. For example, we have studied toxic behavior in online video games using a dataset of millions of incidents. While our findings must certainly be placed in the context of online games, they also have confirmed theories and controlled experiments on the bystander effect and cultural differences in the perception of aggressive and harassing behavior.

Finally, in the past several months a more worrying trend has appeared: organized and politically motivated hate speech and harassment. We have studied this phenomena through the lens of 4chan's Politically Incorrect board, or /pol/. 4chan occupies a somewhat unique position; somewhere just in between the shallow Web and the deep Web. Its anonymous and ephemeral nature has resulted in a community that is responsible for a substantial amount of Internet culture, but also a substantial amount of problems. We examined the usage of hate speech on /pol/ finding it to be significantly higher than on Twitter. We then made a first attempt at measuring raiding behavior. Semi-organized attacks on social web services. Finally, we examine some of the methods of propaganda that this new breed of harassers uses to spread their message and recruit new members.

The end goal of this line of research is to mitigate the damage that these social disruptors exhibit. After a basic understanding of how they work, and potential psychological motivators, we can begin to discuss strategies for dealing with them. Such efforts will require an interdisciplinary approach, with inputs from social scientists as well as computer scientists. Unlike many traditional computer security related questions, we have a further responsibility to design systems to aid victims of these attacks. Although the Web was not designed with these problems in mind, by engaging in cybersafety research, we can continue to ensure the Web remains an open space for all.

### 3.3 Compromised Social Network Accounts – Detection and Incentives

*Manuel Egele (Boston University, US)*

License  Creative Commons BY 3.0 Unported license  
© Manuel Egele

Historically, attackers relied on Sybil or fake accounts to perpetrate their deeds on online social networks. Lately, however, these actors increasingly compromise legitimate accounts to execute their attacks. Compromising legitimate accounts has significant advantages for an attacker. For example, statistical detection approaches that rely on profile characteristics for detection (e.g., profile creation date, distribution of friends or followers) are no longer applicable for detection as these features are perfectly innocuous for benign accounts. Thus, to detect compromised accounts, we developed and present COMPA, a system that models an account's normal behavior over time in a behavioral profile. If a new message posted to an account violates this profile, COMPA considers this as a potential compromise. COMPA then sets out to find a campaign of compromises by identifying accounts that posted similar messages that also violate their accounts' respective behavioral profiles. COMPA successfully detects thousands of campaigns on both, Facebook and Twitter over the duration of three months.

The evaluation of COMPA revealed that Twitter features a prolific ecosystem of so-called follower-markets. These markets are pyramid schemes where volunteer users (victims) allow miscreants to post to their accounts and get additional followers in return. As the attackers control the accounts of their victims, they can sell these accounts as followers to paying customers. In the second part of the talk we shed light on some of these markets and the structures underlying them. Our analysis indicate that the earning potential for attackers who implement such schemes is in the tens of thousands of dollars per month.

### 3.4 Privacy & Safety at Apple


*Julien Freudiger (Apple Computer Inc. – Cupertino, US)*

**License**  Creative Commons BY 3.0 Unported license  
© Julien Freudiger

Protecting user privacy is a core principal in the Apple ecosystem. We discuss a couple of approaches to build great features with privacy, including differential privacy, a technique to learn from users in aggregate while protecting individual user privacy.

### 3.5 Strengthening Weak Identities Through Inter-Domain Trust Transfer

*Oana Goga (MPI-SWS – Saarbrücken, DE)*


**License**  Creative Commons BY 3.0 Unported license  
© Oana Goga  
**Joint work of** G. Venkatadri, C. Zhong, B. Viswanath, N. Sastry, K. Gummadi

On most current websites untrustworthy or spammy identities are easily created. Existing proposals to detect untrustworthy identities rely on reputation signals obtained by observing the activities of identities over time within a single site or domain; thus, there is a time lag before which websites cannot easily distinguish attackers and legitimate users. In this paper, we investigate the feasibility of leveraging information about identities that is aggregated across multiple domains to reason about their trustworthiness. Our key insight is that while honest users naturally maintain identities across multiple domains (where they have proven their trustworthiness and have acquired reputation over time), attackers are discouraged by the additional effort and costs to do the same. We propose a flexible framework to transfer trust between domains that can be implemented in today's systems without significant loss of privacy or significant implementation overheads. We demonstrate the potential for inter-domain trust assessment using extensive data collected from Pinterest, Facebook, and Twitter. Our results show that newer domains such as Pinterest can benefit by transferring trust from more established domains such as Facebook and Twitter by being able to declare more users as likely to be trustworthy much earlier on (approx. one year earlier).



### 3.6 Game BOTs (and economic scale of their black money)

*Huy Kang Kim (Korea University – Seoul, KR)*

License  Creative Commons BY 3.0 Unported license  
© Huy Kang Kim

In the past few years, online games have become popular and have been generating huge profits. As online games become popular and the boundary between virtual and real economies blurs, cheating in games has proliferated in volume and method. Malicious game users do cheating play to level up and accumulate cyber assets in an easy and fast manner without sufficient effort. One of the most widely used tools for cheating in online games is the game bot, which enables users to cheat in a convenient way by automatically performing the required actions. There are also professionally industrialized groups, called as "gold farming group (GFG)", are running numerous machines and runs multiple client programs or game bots to maximize the profit from the online game. They also provide money laundering service to exchange cyber money to real money.

### 3.7 An Army of Me: Sockpuppets in Online Discussion Communities

*Srijan Kumar (Stanford University, US)*

License  Creative Commons BY 3.0 Unported license  
© Srijan Kumar  
Joint work of J. Cheng, J. Leskovec, V. S. Subrahmanian

In online discussion communities, users can interact and share information and opinions on a wide variety of topics. However, some users may create multiple identities, or sockpuppets, and engage in undesired behavior by deceiving others or manipulating discussions. In this work, we study sockpuppetry across nine discussion communities, and show that sockpuppets differ from ordinary users in terms of their posting behavior, linguistic traits, as well as social network structure. Sockpuppets tend to start fewer discussions, write shorter posts, use more personal pronouns such as "I", and have more clustered ego-networks. Further, pairs of sockpuppets controlled by the same individual are more likely to interact on the same discussion at the same time than pairs of ordinary users. Our analysis suggests a taxonomy of deceptive behavior in discussion communities. Pairs of sockpuppets can vary in their deceptiveness, i.e., whether they pretend to be different users, or their supportiveness, i.e., if they support arguments of other sockpuppets controlled by the same user. We apply these findings to a series of prediction tasks, notably, to identify whether a pair of accounts belongs to the same underlying user or not. Altogether, this work presents a data-driven view of deception in online discussion communities and paves the way towards the automatic detection of sockpuppets.

The talk is based on the research paper of the paper with the same title presented at the 26th International World Wide Web Conference, 2017.

### 3.8 The spread of misinformation in social media

*Filippo Menczer (Indiana University – Bloomington, US)*


License  Creative Commons BY 3.0 Unported license  
© Filippo Menczer

Joint work of C. Shao, G. Ciampaglia, O. Varol, A. Flammini

As social media become major channels for the diffusion of news and information, they are also increasingly attractive and targeted for abuse and manipulation. This talk overviews ongoing network analytics, data mining, and modeling efforts to understand the spread of misinformation online and offline. I present machine learning methods to detect astroturf and social bots, as well as theoretical models to study how fake news and fact-checking compete for our collective attention. These efforts will be framed by a case study in which, ironically, our own research became the target of a coordinated disinformation campaign.

### 3.9 Hate Speech: Thoughts on the Role of User Aspects & External Events

*Alexandra Olteanu (IBM TJ Watson Research Center – Yorktown Heights, US)*

License  Creative Commons BY 3.0 Unported license  
© Alexandra Olteanu

The last years have seen an increase in hateful utterances in online social platforms; and this increase has, at times, resulted in harassment and hate crimes on the ground. In this talk, I raise three key questions related to hateful speech online, and present a few preliminary results: (1) Given that there is no universally accepted definition for what constitutes hate speech, how do we evaluate systems dealing with such controversial and subjective concepts? (2) In addition, the lack of a clear definition, can also result in variation in how and when users interpret an online message as hate speech. Thus, another key question is how do user aspects impact the way in which they perceive online hateful chatter? (3) Finally, empirical evidence suggest that external events often trigger an increase in hateful speech online. How can we quantify the impact of external events on the prevalence and type of hateful chatter online?

### 3.10 Your words betray you! The role of language in cyber crime investigations

*Awais Rashid (Lancaster University, GB)*

License  Creative Commons BY 3.0 Unported license  
© Awais Rashid

Online social media and networks are increasingly utilized in cyber criminal activities. Sophisticated criminals often take steps to avoid revealing critical information about themselves or their activities. This poses significant challenges for legitimate law enforcement activity to protect victims and apprehend criminals. In this talk I will reflect on experiences in two large-scale projects and discuss the challenges of analyzing online activities of cyber criminals, including deception and the use of specialized vocabulary to share illegal sexual materials.

I will then highlight how advances in computational analysis of natural language can help overcome these challenges. Both projects have seen real-world deployments, so the talk will cover both scientific value of linguistic analysis in this context and insights from practical experiences in law enforcement settings. I will conclude by discussing the implications for cyber safety research and the need for a shift – to a multi-disciplinary approach for tackling cyber crime.

### 3.11 Combating Friend Spam Using Social Rejections

*Michael Sirivianos (Cyprus University of Technology – Lemesos, CY)*

**License** © Creative Commons BY 3.0 Unported license  
© Michael Sirivianos

**Joint work of** Q. Cao, X. Yang, K. Munagala

**Main reference** Qiang Cao, Michael Sirivianos, Xiaowei Yang, Kamesh Munagala: “Combating Friend Spam Using Social Rejections”, in Proc. of the 35th IEEE International Conference on Distributed Computing Systems, ICDCS 2015, Columbus, OH, USA, June 29 - July 2, 2015, pp. 235–244, IEEE Computer Society, 2015.

**URL** <http://dx.doi.org/10.1109/ICDCS.2015.32>

Unwanted friend requests in online social networks (OSNs), also known as friend spam, are among the most evasive malicious activities. Friend spam can result in OSN links that do not correspond to social relationship among users, thus pollute the underlying social graph upon which core OSN functionalities are built, including social search engine, ad targeting, and OSN defense systems. To effectively detect the fake accounts that act as friend spammers, we propose a system called Rejecto. It stems from the observation on social rejections in OSNs, i.e., even well maintained fake accounts inevitably have their friend requests rejected or they are reported by legitimate users. Our key insight is to partition the social graph into two regions such that the aggregate acceptance rate of friend requests from one region to the other is minimized. This design leads to reliable detection of a region that comprises friend spammers, regardless of the request collusion among the spammers. Meanwhile, it is resilient to other strategic manipulations. To efficiently obtain the graph cut, we extend the Kernighan-Lin heuristic and use it to iteratively detect the fake accounts that send out friend spam. Our evaluation shows that Rejecto can discern friend spammers under a broad range of scenarios and that it is computationally practical.

### 3.12 Cyberfraud and Scams

*Gianluca Stringhini (University College London, GB)*

**License** © Creative Commons BY 3.0 Unported license  
© Gianluca Stringhini

Fraud perpetrated by humans (rather than bots) is becoming an increasing problem on online services. Due to its nature, the techniques designed to identify malicious activity are not enough to detect and mitigate it. This talk covers the problems of online dating scams and identity theft, highlighting some research challenges and findings coming from our recent research.

### 3.13 The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources

*Savvas Zannettou (Cyprus University of Technology – Lemesos, CY)*

**License** © Creative Commons BY 3.0 Unported license  
© Savvas Zannettou

**Joint work of** T. Caulfield, E. De Cristofaro, N. Kourtellis, I. Leontiadis, M. Sirivianos, G. Stringhini, J. Blackburn

**Main reference** Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, Jeremy Blackburn: “The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources”, in Proc. of the 2017 Internet Measurement Conference, IMC 2017, London, United Kingdom, November 1-3, 2017, pp. 405–417, ACM, 2017.

**URL** <http://dx.doi.org/10.1145/3131365.3131390>

As the number and diversity of news sources on the Web grows, so does the opportunity for alternative sources of information production. The emergence of mainstream social networks like Twitter and Facebook makes it easier for misleading, false, and agenda driven information to quickly and seamlessly spread online, deceiving people or influencing their opinions. Moreover, the increased engagement of tightly knit communities, such as Reddit and 4chan, compounds the problem as their users initiate and propagate alternative information not only within their own communities, but also to other communities and social media platforms across the Web. These platforms thus constitute an important piece of the modern information ecosystem which, alas, has not been studied as a whole. In this paper, we begin to fill this gap by studying mainstream and alternative news shared on Twitter, Reddit, and 4chan. By analyzing millions of posts around a variety of axes, we measure how mainstream and alternative news flow between these platforms. Our results indicate that alt-right communities within 4chan and Reddit can have a surprising level of influence on Twitter, providing evidence that “fringe” communities may often be succeeding in spreading these alternative news sources to mainstream social networks and the greater Web.

## 4 Working groups

### 4.1 Algorithmic Biases

*Seda F. Guerses (KU Leuven, BE) and Savvas Zannettou (Cyprus University of Technology – Lemesos, CY)*

**License** © Creative Commons BY 3.0 Unported license  
© Seda F. Guerses and Savvas Zannettou

This working group focused on various biases that exist in research, ranging from biases in data, annotations, algorithms and systems.

Some examples include:


- **Biases in Dataset.** The use of snowball sampling for data acquisition eventually leads to datasets with biases, as one is more likely to get high degree nodes.
- **Biases in Annotation.** In many cases, we employ manual annotators to label our dataset. However, there are cases where the annotations are not trivial and people may not agree. This fact may introduce un-wanted biases in our dataset during the annotation phase.

- **Algorithmic Biases.** Some decision making algorithms have been shown to discriminate on the basis of race and other protected categories. For example, an example used in the US criminal justice was shown to lead to increased criminalization and imprisonment of African American men.

During the discussions, the participants realized that bias can exist in multiple forms and there are a lot of challenges when defining bias. Therefore, as an open research direction, we discussed possible definitions of biases, and if tools can be implemented that will aim to detect biases in algorithms and/or systems. Also, evaluating the percentage and impact of biases on systems is not a straightforward task, as this depends on the use case as well as the criticality of the used data.

## 4.2 Ethics

*Seda F. Guerses (KU Leuven, BE) and Savvas Zannettou (Cyprus University of Technology – Lemesos, CY)*

License  Creative Commons BY 3.0 Unported license  
© Seda F. Guerses and Savvas Zannettou

This working group focused on the discussion of ethical considerations when doing research. We started from observing a lack of general guidelines with regard to the ethical aspects of research and usually this comes to the Institutional Review Board (IRB) of the involved institutions. Also, we noted that there are a lot of risks when doing cybersecurity experiments with real users.

In such cases we should make sure that we enforce subject integrity, i.e., making users aware of the underlying risks. Also, in the case of interdisciplinary studies, there is a need of close communication between the involved researchers, as ethical issues may arise in one of the multiple dimensions of the study.

As the main research direction for this Working Group, we proposed the compilation of general ethics guidelines that can assist researchers in understanding on whether their research ideas are ethical or not and what actions should be done to make sure of the legality of the research.

## 4.3 Cyberfraud

*Savvas Zannettou (Cyprus University of Technology – Lemesos, CY)*

License  Creative Commons BY 3.0 Unported license  
© Savvas Zannettou

This working group focused on the problem of cyberfraud. We acknowledged that cyberfraud has evolved the last few years, as we observed events of political distraction as well as the generation of fake news. To this end, we identified the following research problems:

1. How do all scams (e.g., buying likes, etc) relate to the interference of election?
2. Evaluating approaches can be difficult, as one method can mitigate the problem on one Web community but users can go to a different community, hence the problem persists.
3. What make political actors thrive? There is a need to identify the techniques that these actor use.

4. How to identify malicious intentions from benign ones?
5. Automated detection tools will generate false positives, which is something that needs to be carefully addressed as users might lose trust in the Web platform.

As the main research direction for this Working Group, we proposed the compilation of a cyberfraud taxonomy where we will cover the whole spectrum of cyberfraud. Specifically, it should contain the various tools exploited by malicious users (e.g., bots), the motives behind the act of scamming users online as well as the various involved actors, and the various types of fraudulent activities.

#### 4.4 Hate Speech

*Savvas Zannettou (Cyprus University of Technology – Lemesos, CY)*

License  Creative Commons BY 3.0 Unported license  
© Savvas Zannettou

This working group focused on the definition and possible research directions for the emerging problem of hate speech in online Web communities. We observed that there is no formal definition for hate speech, therefore, we proposed that a formal definition should include the bias of the offender as well as his intentions, the targeted group (based on its traits), and the victim's perspective.

Nevertheless, there are a lot of open questions regarding hate speech:

1. When does a post qualify to be hate speech? Only if is in public or in private too?
2. How to identify the intentions of the offender? For instance, a post might be treated as a joke from the offender's perspective.
3. If we manage to detect hate speech how we mitigate the problem?

For the latter, we noted that silencing is not a viable solution, as users will lose trust in the Web community. Therefore, we proposed the implementation of solutions that aim to counter the speech rather than stopping it, decrease the visibility of the hateful content, create a support service to help the victim overcome the effect from the insults, and develop containment systems that will stop the diffusion process within the Web community.

The main tangible research directions that were discussed during this working group include:

1. Understanding the effects of different definition or detection mechanisms for hate speech;
2. Evaluating strategies after the detection of hate speech;
3. Contagion experiments on Web communities, e.g. Twitter and the followers network; and
4. Comparing multiple intervention mechanisms as well as tackling hate speech in multiple languages.

## Participants

- Zinaida Benenson  
Universität Erlangen-  
Nürnberg, DE
- Jeremy Blackburn  
University of Alabama at  
Birmingham, US
- Emiliano De Cristofaro  
University College London, GB
- Julien Dreux  
Facebook – London, GB
- Manuel Egele  
Boston University, US
- Julien Freudiger  
Apple Computer Inc. –  
Cupertino, US
- Oana Goga  
MPL-SWS – Saarbrücken, DE
- Seda F. Gürses  
KU Leuven, BE
- Huy Kang Kim  
Korea University – Seoul, KR
- Christiane Kuhn  
TU Dresden, DE
- Srijan Kumar  
Stanford University, US
- Ilias Leontiadis  
Telefónica Research –  
Barcelona, ES
- Filippo Menczer  
Indiana University –  
Bloomington, US
- Prateek Mittal  
Princeton University, US
- Alexandra Olteanu  
IBM TJ Watson Research Center  
– Yorktown Heights, US
- Awais Rashid  
Lancaster University, GB
- Ahmad-Reza Sadeghi  
TU Darmstadt, DE
- Stefan Schiffner  
ENISA – Athens, GR
- Michael Sirivianos  
Cyprus University of Technology  
– Lemesos, CY
- Gianluca Stringhini  
University College London, GB
- Thorsten Strufe  
TU Dresden, DE
- Savvas Zannettou  
Cyprus University of Technology  
– Lemesos, CY

