Integrity 2021: Integrity in Social Networks and Media

Lluís Garcia-Pueyo
Facebook
Menlo Park, USA
lgp@fb.com

Anand Bhaskar
Facebook
Menlo Park, USA
anandb@fb.com

Roelof van Zwol
Pinterest
San Francisco, USA
roelof@pinterest.com

Timos Sellis
Facebook
Menlo Park, USA
tsellis@fb.com

Gireeja Ranade
UC Berkeley
Berkeley, USA
ranade@eecs.berkeley.edu

Prathyusha Senthil Kumar
Facebook
Menlo Park, USA
prathyushas@fb.com

Yu Sun
Twitter
San Francisco, USA
ysun@twitter.com

Joy Zhang
Airbnb
San Francisco, USA
joycmu@gmail.com

ABSTRACT

This is the proposal for the second edition of the Workshop on Integrity in Social Networks and Media, Integrity 2021, following the success of the first Workshop held in conjunction with the 13th ACM Conference on Web Search and Data Mining (WSDM) in Houston, Texas, USA [3]. The goal of the workshop is to bring together researchers and practitioners to discuss content and interaction integrity challenges in social networks and social media platforms. The event consists on (1) a series of invited talks by reputed members of the Integrity community from both academia and industry, (2) a call-for-papers for contributed talks, and (3) a panel with the speakers.

KEYWORDS

social networks, social media, integrity, quality, misinformation, fairness

ACM Reference Format:
Lluís Garcia-Pueyo, Anand Bhaskar, Roelof van Zwol, Timos Sellis, Gireeja Ranade, Prathyusha Senthil Kumar, Yu Sun, and Joy Zhang. 2020. Integrity 2021: Integrity in Social Networks and Media. In *Proceedings of WSDM '21.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

WSDM '21, March 08–12, 2021, Jerusalem, Israel

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 WORKSHOP DESCRIPTION

Integrity 2021 aims to repeat the success achieved in the previous edition, Integrity 2020, which was hosted within WSDM'20 at Houston, TX. In the previous edition [3], 4 industry leads (from Facebook, Twitter, Pinterest, and Airbnb) and 3 academic experts (from UC Berkeley, MIT, and Facebook Core-Data-Science) exposed challenges, solutions, and ongoing research in areas such as Misinformation, Bias in Machine Learning models, Content-based detecton, Display Advertising Integrity, Behavioral-based detection, and others. The previous workshop resulted in fruitful discussions and engagement from the audience, and a unanimous push towards organizing a recurring workshop exploring these problems and potential solutions, with participation from academics and industry researchers. Besides, there is a strong interest in the community in integrity, with several related workshops and conferences on related topics [1, 2, 4].

In the past decade, social networks and social media sites, such as Facebook and Twitter, have become the default channels of communication and information. The popularity of these online portals has exposed a collection of integrity issues: cases where the content produced and exchanged compromises the quality, operation, and eventually the integrity of the platform. Examples include misinformation, low quality and abusive content and behaviors, and polarization and opinion extremism. There is an urgent need to detect and mitigate the effects of these integrity issues, in a timely, efficient, and unbiased manner.

This workshop aims to bring together top researchers and practitioners from academia and industry, to engage in a discussion about algorithmic and system aspects of integrity challenges. The WSDM Conference, that combines Data Mining and Machine Learning with research on Web and Information Retrieval offers the ideal forum for such a discussion, and we expect the workshop to be of interest to everyone in the community. The topic of the workshop is also interdisciplinary, as it overlaps with psychology, sociology, and economics, while also raising legal and ethical questions, so we expect it to attract a broader audience.

As indicated by the organizing committee and the speaker list, the workshop aims to bring together researchers and practitioners

Table 1: Workshop schedule

Time	Event
8:45	Welcoming remarks
9:00	Invited Talk - **Miklos Racz** (Assistant Professor, Princeton University)
9:45	Invited Talk - **Panagiotis Papadimitriou** (Facebook, Director of Engineering at News Feed Integrity)
10:30	Coffee Break
11:00	Invited Talk - **Jeremy Blackburn** (Assistant Professor, Binghamton University)
11:45	Invited Talk - **Ding Zhou** (Snap, Senior Director of Engineering for Discovery and Content)
12:30	Lunch
1:30	Invited Talk - **Bruno Ordozgoiti** (Postdoctoral research, Aalto University)
2:15	Invited Talk - **Grace Tang** (Sr. Staff Machine Learning Engineer (Anti-Abuse) at LinkedIn)
3:00	Coffee Break
3:30	Invited Talk - **Keshava Subramanya** (Ads Intelligence Lead at Pinterest)
4:15	Invited Talk - **Axel Bruns** (Professor at Queensland University of Technology (QUT))
5:00	Speakers Panel & Closing remarks

from both industry and academia, leading to exchange of knowledge and cross-cutting collaborations.

Workshop topics:

- **Low quality, borderline, and offensive content and behaviors:** Methods for detecting and mitigating low quality and offensive content and behaviors, such as click bait, fake engagement, nudity and violence, bullying, and hate speech.
- **Personalized treatment of low quality content:** Identification, measurement and reduction of bad experiences.
- **Misinformation:** Detecting and combating misinformation; Deep and shallow fakes; Prevalence and virality of misinformation; Misinformation sources and origins; Source and content credibility.
- **Integrity in Polarization:** Models and metrics for polarization; Echo chambers and filter bubbles; Opinion Extremism and radicalization; Algorithms for mitigating polarization.
- **Fairness in Integrity:** Ensure fairness in the detection and mitigation of integrity issues with respect to sensitive attributes such as gender, race, sexual orientation, and political affiliation.

2 WORKSHOP FORMAT

- **Duration & Format:** Full day, invited-speakers.
- **Number of participants (estimated):** 40
- **Schedule (tentative):** Table 1 details the proposed schedule. Note that the speakers in the schedule have **confirmed** their participation, and their bios and tentative titles for their talks are provided in this proposal.
- **Call-for-papers, Poster session and Contributed talks:** The committee will open the workshop to additional contributors through contributed talks and a poster session, via a call-for-papers. Schedule will be adapted to allocate contributed events.

3 CONFIRMED INVITED SPEAKERS

3.1 Miklos Racz (Princeton University)

Miklos Z. Racz is an assistant professor at Princeton University in the ORFE department, as well as an affiliated faculty member at the Center for Statistics and Machine Learning (CSML). Before coming to Princeton, he received his PhD in Statistics from UC Berkeley and was then a postdoc in the Theory Group at Microsoft Research, Redmond. Miki's research focuses on probability, statistics, and their applications, and he is particularly interested in network science and the spread of (mis/dis)information. Miki's research and teaching has been recognized by Princeton's Howard B. Wentz, Jr. Junior Faculty Award, a Princeton SEAS Innovation Award, and an Excellence in Teaching Award.

Title: An Adversarial Perspective on Network Disruption

Abstract: I will discuss a simple new model of network disruption, where an adversary can take over a limited number of user profiles in a social network with the aim of maximizing disagreement and/or polarization in the network. I will present both theoretical and empirical results. Theoretically, we characterize aspects of the adversary's optimal decisions and prove bounds on their disruptive power. Furthermore, we present a detailed empirical study of several natural algorithms for the adversary on both synthetic networks and real world (Reddit and Twitter) data sets. These show that even simple, unsophisticated heuristics, such as targeting centrists, can disrupt a network effectively. This is based on joint work with Mayee F. Chen.

3.2 Panagiotis Papadimitriou (Facebook)

Panagiotis Papadimitriou, is the Connection Integrity Eng Pillar Lead at Facebook. His team is responsible for the Integrity of News Feed, Stories, Pages and Facebook App Monetization products. Prior to Facebook, Panagiotis was the Senior Director of Data Science and Engineering team at Upwork, the world's largest online workplace. At Upwork, Panagiotis built and ran the teams responsible for job search, applicant recommendations and ML & experimentation infra. Panagiotis received his PhD and MS degrees from Stanford University and a BS from National Technical University of Athens.

During his graduate studies he received a scholarship from the Onassis Foundation, Yahoo and Stanford University.

Title: Optimizing for people's safety at Facebook App while protecting freedom of expression.

3.3 Ding Zhou (Snap)

Ding Zhou is the senior director of Engineering for Content and Discovery at Snap[1]. Previously he led the Ads Engineering team at Pinterest, and was the VP of engineering at Doordash. Trust and Safety at Snap is a cornerstone of the experience the Snap brings to millions of users to improve the way they live and communicate.

Title: Trust and Safety at Snap

3.4 Jeremy Blackburn (Binghamton University)

Jeremy Blackburn is an Assistant Professor in the Department of Computer Science at Binghamton University and co-founder of the International Data-driven Research for Advanced Modeling and Analysis Lab (iDRAMA Lab). Jeremy is broadly interested in data science, with a focus on large-scale measurements and modeling. His largest line of work is in understanding jerks on the Internet. His award winning research into understanding toxic behavior, hate speech, and fringev and extremist Web communities has been covered in the press by The Washington Post, the New York Times, The Atlantic, The Wall Street Journal, the BBC, and New Scientist, among others. Prior to his appointment at Binghamton University, Jeremy was an Assistant Professor in the Department of Computer Science at the University of Alabama at Birmingham. Prior to that, Jeremy was an Associate Researcher at Telefonica Research in Barcelona, Spain. Jeremy also has a finite Erdos-Bacon number of O(7).

Title: WTF is Going On?!?! A 5 Year Retrospective of Data-driven Research on Assholes. Towards a Data-driven Approach to Understanding Online Extremism

Abstract: In this talk, I present our efforts over the past 5+ years to measure, model, understand, and mitigate abusive and dangerous online behavior. I will start by discussing some of my earliest work that made use of large-scale data to automate the detection of toxic behavior in online video games and the implications it has for understanding socio-technical issues. Next, I will demonstrate how relatively small, fringe Web communities have outsized influence on the greater Web in terms of spreading information, as well as making coordinated attacks against other communities. Finally, I will present our study of the evolution of the Manosphere, which provides large-scale, longitudinal evidence of what amounts to a radicalization pipeline.

3.5 Bruno Ordozgoiti (Aalto University)

Bruno Ordozgoiti is a postdoctoral researcher at Aalto University. His recent work is motivated by the problem of polarized behavior in social media, focusing chiefly on the detection of conflicting structures in signed networks, but also introducing notions of polarization into fundamental computational problems like clustering. Other contributions of his range from kernel methods to robust matrix factorization, and have been published at some of the leading international data mining venues (WWW, CIKM, ICDM, ECML-PKDD). He earned his PhD in 2018 at Universidad Politécnica de Madrid.

Title: Detecting polarized structures in social media.

Abstract: Over the last few years, social media platforms have become a key channel through which news outlets and political leaders communicate with their audiences. and the contention for the public's attention now takes place in an overpopulated, highly competitive arena. This incentivizes the use of sensationalistic headlines, clickbait and compelling memes to secure user engagement, usually favoring polarizing narratives instead of thoughtful, nuanced and well-researched news pieces. Rather than promoting healthy debate, these practices are arguably inciting confrontational, toxic and abusive interactions online. Can we use computational methods to detect and mitigate this type of behavior?

3.6 Grace Tang (Anti-Abuse at Linkedin)

Grace Tang is a Senior Staff Machine Learning Engineer on the Trust AI Team at LinkedIn. She has worked on combating fake accounts, scraping, harassment, job fraud, and other abuses. Currently, she works across abuse domains, focusing on integrating detection systems together to achieve defense in depth.

Title: Integrity Ecosystem at LinkedIn

3.7 Axel Bruns (Queensland University of Technology)

Prof. Axel Bruns is a Professor in the Digital Media Research Centre at Queensland University of Technology in Brisbane, Australia, and a Chief Investigator in the ARC Centre of Excellence for Automated Decision-Making and Society. His books include Are Filter Bubbles Real? (2019) and Gatewatching and News Curation: Journalism, Social Media, and the Public Sphere (2018), and the edited collections Digitizing Democracy (2019), the Routledge Companion to Social Media and Politics (2016), and Twitter and Society (2014). His current work focusses on the study of user participation in social media spaces such as Twitter, and its implications for our understanding of the contemporary public sphere, drawing especially on innovative new methods for analysing 'big social data'. He served as President of the Association of Internet Researchers in 2017–19.

Title: Social Media and the News: Approaches to the Spread of (Mis)information

Abstract: This paper presents an overview of several research initiatives in the ARC Centre of Excellence for Automated Decision-Making and Society and QUT Digital Media Reseaech Centre that examine the spread of information, misinformation, and disinformation across social media platforms especially in the context of the COVID-19 pandemic. Drawing on a variety of methods from large-scale analytics to detailed forensic analysis, we examine the differences in the dissemination dynamics of mainstream and fringe news content; trace the spread of conspiracy theories and other coronavirus misinformation to identify the key points of inflection and amplification; explore methods for the detection of coordinated inauthentic behaviour; and examine the impact of automated content moderation.

[1] https://www.snap.com/en-US/

3.8 Keshava Subramanya (Pinterest)

Keshava Subramanya leads the Ads Intelligence efforts at Pinterest. This effort includes optimizing ad campaigns throughout the ad life-cycle with pre-setup, post-setup ad review efforts and finally ads delivery optimization and advertiser recommendations. Keshava loves building distributed and large scale systems and has previously worked on recommender systems at Netflix and on Bing Search at Microsoft. Keshava received his Masters from the University of California at Santa Barbara.

Title: Ads Integrity at Pinterest

4 ORGANIZERS

Lluis Garcia-Pueyo, Facebook, is an Engineering Manager at Facebook, leads the News Feed Integrity Distribution pillar focussing on personalization, discovery and reduction of negative experiences in News Feed and Stories ranking. Prior to this, he worked in information extraction and information retrieval at Google Research, and multimedia retrieval and display advertising at Yahoo Research. Lluis holds an MS in Computer Science from the Universitat Politècnica de Catalunya. His research has been published in top-tier conferences such as WWW, KDD, SIGIR, ACM Multimedia, and WSDM, and he is a member of the PC for KDD, WWW and other conferences. He is an organizer of the internal Integrity Week conference at Facebook, which hosts 200+ research scientists, data scientists and engineers in Social Network Integrity topics. He organized the Integrity 2020 edition of this workshop at WSDM'20.

Anand Bhaskar, Facebook. Anand Bhaskar is a Research Scientist at Facebook, where he works on building and incorporating content quality signals into News Feed ranking and studying the network effects of ranking changes. Prior to that, he was a postdoctoral researcher at Stanford University and HHMI, where he applied techniques from statistics, computer science, and applied mathematics to large-scale genomic datasets for addressing scientific questions such as the genetic basis of disease, human demographic history, and forensics, among others. His work has been published in journals such as PNAS and the Annals of Statistics, and conferences such as VLDB and SIGMOD. He received a Ph.D. in Computer Science and an M.A. in Statistics from the University of California, Berkeley, and M.Eng. and B.S. degrees in Computer Science from Cornell University. His research has been supported by a Berkeley Fellowship, Simons-Berkeley Research Fellowship, Japan Society for the Promotion of Science Postdoctoral Fellowship, and Stanford CEHG Postdoctoral Fellowship. He organized the Integrity 2020 edition of this workshop at WSDM'20.

Roelof van Zwol is the head of Ads Quality at Pinterest. The team is responsible for (1) helping advertisers define the audience they want to reach through services such as Act-alike modeling, interest targeting, etc, as well as the ML models that power the ads delivery system to determine which ads to show to a Pinner in a generalized second price auction. Previously, Roelof was the Director of Product Innovation at Netflix. There he was responsible for the innovation of Netflix's content promotion and acquisition algorithms. Prior to joining Netflix, Roelof managed the multimedia research team at Yahoo!, first from Barcelona, Spain, and later from Yahoo!'s headquarters in California. He started his career in academia as an assistant professor in the Computer Science department in Utrecht, the Netherlands, after finishing his PhD at the University of Twente in Enschede, the Netherlands.

Timos Sellis is a visiting scientist at Facebook and an Adjunct Professor in Computer Science at Swinburne University of Technology, where he also served as the director of the Data Science Research Institute (2026-20). He received his M.Sc. degree from Harvard University (1983) and Ph.D. degree from the University of California at Berkeley (1986). He has served as a professor at the University of Maryland (1986-92), the National Technical University of Athens (1992-2013), and was the inaugural Director of the Institute for the Management of Information Systems of the "Athena" Research Center (2007-13).He is IEEE Fellow (2009) and an ACM Fellow (2013), for his contributions to database systems and data management. In March 2018 he received the IEEE TCDE Impact Award, for contributions to database systems research and broadening the reach of data engineering research.

Gireeja Ranade is Assistant Teaching Professor in EECS at UC Berkeley. Before joining the faculty at UC Berkeley, Dr. Ranade was a Researcher at Microsoft Research AI in the Adaptive Systems and Interaction Group. She also designed and taught the first offering for the new course sequence EECS16A and EECS16B in the EECS department at UC Berkeley and received the 2017 UC Berkeley Electrical Engineering Award for Outstanding Teaching. Dr. Ranade received her PhD in Electrical Engineering and Computer Science from the University of California, Berkeley, and her undergraduate degree from MIT in Cambridge, MA. Dr. Ranade's work on Understanding Misinformation in recent years is specially relevant to the Integrity Workshop.

Prathyusha Senthil Kumar, Facebook, is an Engineering Manager at Facebook, where she leads News Feed Integrity efforts leveraging machine learning techniques to understand and utilize content quality in feed ranking and to reduce subjective bad experiences through personalized ranking interventions. Prior to joining Facebook, she led the Search Query Understanding and Relevance Ranking applied research teams at Ebay Inc. Prior to that she worked as a researcher at EBay applying machine learning, natural language processing and information retrieval for various search problems. She holds a master's degree in Computer Science from the University of Texas at Austin and an undergraduate degree in Information Technology from the College of Engineering Guindy, Anna University. Her research has been published in conferences such as SIGIR, CIKM and IEEE.

Yu Sun, Twitter, is a Machine Learning Engineer at TwitterâĂŹs Search team, which builds one of the most popular real time search engines for tweets and users. His research interests include context-aware recommendation and personalization. His work has been published in top tier venues such as ICML, NeurIPS, KDD, and WWW. He obtained his PhD in computer science from the University of Mel- bourne in 2017 and he received the Best Student Paper Award for the Applied Data Science track in KDD 2016.

Joy Zhang is the Head of AI Labs at Airbnb, leading the effort of building AI technologies for future travel experiences. Dr. Zhang received his Ph.D. from Carnegie Mellon University in 2008 and joint the faculty of Carnegie Mellon University after graduation. He received 18 grants of total $22M from NSF, DARPA, Army Research Office, Northrop Grumman, Nokia, Google, Intel and Yahoo!. in the area of statistical natural language processing, mobile computing and user behavior modeling. In 2013, he joined Facebook following Facebook's acquisition of the startup company Jibbigo which Dr. Zhang was one of the founding members. Dr. Zhang built the machine translation team at Facebook and built the Natural Language Understanding team two years later. At Applied Machine Learning (AML), NLU team developed DeepText, a deep learning system for text understanding which is powering most Facebook's NLP features such as spam detection, suicide prevention, hate speech detection. From 2017 to 2018, he managed the Content Quality Modeling team of News Feed Integrity building ML systems to battle low quality content on Facebook such as click-bait, engagement-bait and fake news.

5 RELATED WORKSHOPS

5.1 Integrity 2020

Hosted in WSDM 2020 at Houston, TX[2] , the previous edition of this workshop [3] brought together Integrity experts from industry leaders with researchers, and focussed on content-based integrity, integrity and abuse in display advertising, misinformation, behavioral analysis, and integrity challenges for machine learning applications.

5.2 CyberSafety 2019

Hosted In The Web Conference 2019[3], this workshop focussed on anomalous behaviors such as fraudulent engagement, misinformation and propaganda, user deception and scams, harassment, hate speech, cyberthreats, cyberbullying on social networks.

5.3 MisinfoWorkshop2019

The International Workshop on Misinformation, Computational Fact-Checking and Credible Web, in The Web Conference 2019[4], focussed on computational methodology for Misinformation and fact-checking detection, and Ethical pitfalls and solutions, as well as Education on Misinformation.

5.4 FATES on the Web'19

Hosted by The Web Conference 2019, this workshop, with a focus on Social Sciences, discussed issues such as Transparency, Credibility, Fairness, Bias and Ethics in computational research and analysis.

5.5 OCeANS Workshop'18

The Opinions, Conflict, and Abuse in a Networked Society[5], hosted in ACM SIGKDD'18, had talks on crowdsourcing and the effects of the usage of user data in detection tasks, methods for low-quality content detection, and adversarial system design.

5.6 ROME 2019

The Workshop on Reducing Online Misinformation Exposure[6], colocated with SIGIR 2019, presented work on subjectivity on crowdsourcing, credibility and bias, medical misinformation, user-generated video verification and time-sensitive fact-checking.

5.7 Truth Discovery 2019

The Truth Discovery and Fact Checking: Theory and Practice[7] workshop, collocated with ACM SIGKDD'19, discussed a broad range of topics related to the Misinformation discovery problem: general architectures, natural language processing for claims, using fact-checking in supervised settings, information extraction, plagiarism.

REFERENCES

[1] Aws Albarghouthi and Samuel Vinitsky. 2019. Fairness-Aware Programming. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 211–219. https://doi.org/10.1145/3287560.3287588

[2] Guillaume Bouchard, Guido Caldarelli, and Vassilis Plachouras. 2019. ROME 2019: Workshop on Reducing Online Misinformation Exposure. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1426–1428. https://doi.org/10.1145/3331184.3331645

[3] Lluis Garcia-Pueyo, Anand Bhaskar, Panayiotis Tsaparas, Aristides Gionis, Tina Eliassi-Rad, Maria Daltayanni, Yu Sun, and Panagiotis Papadimitriou. 2020. Integrity 2020: Integrity in Social Networks and Media. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (Houston, TX, USA) *(WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 905–906. https://doi.org/10.1145/3336191.3371880

[4] Richard Han and Neil Shah. 2019. Cybersafety 2019: The 4th Workshop on Computational Methods in Online Misbehavior. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 146–147. https://doi.org/10.1145/3308560.3316493

[2] http://integrity-workshop.org/
[3] https://cybersafety2019.github.io/
[4] https://sites.google.com/view/misinfoworkshop
[5] https://sites.google.com/view/oceans-kdd2018/home
[6] https://rome2019.github.io/
[7] https://truth-discovery-kdd2019.github.io/